January 2013

Multiple Calibrations in Integrative Data Analysis: A Simulation Study and Application to Multidimensional Family Therapy

Kristin Wynn Hall
University of South Florida, Khall2@health.usf.edu

Multiple Calibrations in Integrative Data Analysis:

A Simulation Study and Application to Multidimensional Family Therapy

by

Kristin Hall

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Public Health
Department of Epidemiology and Biostatistics
College of Public Health
University of South Florida

Major Professor: Getachew Dagne, Ph. D.
Co-Major Professor: Wei Wang, Ph. D.
Paul Greenbaum, Ph. D.

Date of Approval:
June 27, 2013

Keywords: pooled data analysis, longitudinal data analysis, moderated nonlinear factor analysis, multiple imputation, latent variables, commensurate measures, adolescent substance use

Acknowledgments

Foremost, I would like to thank my committee members, Dr. Dagne, Dr. Wang and Dr. Greenbaum, for the immense amount of valuable advice and guidance throughout the entire learning process of a Master's Thesis. I would particularly like to express my gratitude to Dr. Wang who appointed me to a Graduate Assistant position last year which not only introduced me to this topic but provided a path for my curiosity in statistics.

I would also like to acknowledge Craig Henderson, Lisa Kan, Anders Alexandersson, Howard Liddle, Gayle Dakof and the support funded by a grant from The National Institute on Drug Abuse R01 DA029089 (PIs: Greenbaum & Henderson) that made this project possible.

Furthermore, I would like to express my deepest appreciation to my parents, Sterling and Pam Hall, and grandparents, Charles and Joanne Hall for believing in me, and for the tremendous amount of support and encouragement throughout my entire education.

Table of Contents

List of Tables

List of Figures

Abstract

A recent advancement in statistical methodology, Integrative Data Analyses (IDA Curran & Hussong, 2009) has led researchers to employ a calibration technique as to not violate an independence assumption. This technique uses a randomly selected, simplified correlational structured subset, or calibration, of a whole data set in a preliminary stage of analysis. However, a single calibration estimator suffers from instability, low precision and loss of power. To overcome this limitation, a multiple calibration (MC; Greenbaum et al., 2013; Wang et al., 2013) approach has been developed to produce better estimators, while still removing a level of dependency in the data as to not violate independence assumption. The MC method is conceptually similar to multiple imputation (MI; Rubin, 1987; Schafer, 1997), so MI estimators were borrowed for comparison.

A simulation study was conducted to compare the MC and MI estimators, as well as to evaluate the performance of the operating characteristics of the methods in a cross classified data characteristic design. The estimators were tested in the context of assessing change over time in a longitudinal data set. Multiple calibrations consisting of a single measurement occasion per subject were drawn from a repeated measures data set, analyzed separately, and then combined by the rules set forth by each method to produce the final results. The data characteristics investigated were effect size, sample size, and the number of repeated measures per subject. Additionally, a real data application of an MC approach in an IDA framework was conducted on data from three completed,

randomized controlled trials studying the treatment effects of Multidimensional Family Therapy (MDFT; Liddle et al., 2002) on substance use trajectories for adolescents at a one year follow-up.

The simulation study provided empirical evidence of how the MC method preforms, as well as how it compares to the MI method in a total of 27 hypothetical scenarios. There were strong asymptotic tendencies observed for the bias, standard error, mean square error and relative efficiency of an MC estimator to approach the whole set estimators as the number of calibrations approached 100. The MI combination rules proved not appropriate to borrow for the MC case because the standard error formulas were too conservative and performance with respect to power was not robust. As a general suggestion, 5 calibrations are sufficient to produce an estimator with about half the bias of a single calibration estimator and at least some indication of significance, while 20 calibrations are ideal. After 20 calibrations, the contribution of an additional calibration to the combined estimator greatly diminished.

The MDFT application demonstrated a successful implementation of 5 calibration approach in an IDA on real data, as well as the risk of missing treatment effects when analysis is limited to a single calibration's results. Additionally, results from the application provided evidence that MDFT interventions reduced the trajectories of substance use involvement at a 1-year follow-up to a greater extent than any of the active control treatment groups, overall and across all gender and ethnicity subgroups. This paper will aid researchers interested in employing a MC approach in an IDA framework or whenever a level of dependency in a data set needs to be removed for an independence assumption to hold.

Chapter 1: Introduction

In the context of this paper, a calibration is referred to as a randomly selected subset of a single measurement occasion per subject that is drawn from a data set containing repeated measures per subject. Using a calibration as a simplified correlational structured subset representation of the whole set has been suggested for the purpose of developing a latent variable measurement model in the first stage of an Integrative Data Analysis (IDA; Bauer & Hussong, 2009; Curran et al., 2008; Curran & Hussong, 2009; Hussong et al., 2013) conducted on longitudinal data. More specifically, a calibration estimates a Moderated Nonlinear Factor Analysis (MNLFA; Bauer & Hussong, 2009) model that conceptualizes the underlying construct (primary outcome) of interest as a latent variable, and uses multiple indicators (outcome measures) simultaneously to generate factor scores for the full set of observations, on which a subsequent (second stage) longitudinal analysis can be conducted (Henderson et al., 2013). The two stage solution was necessary because existing software could not both accommodate the complexity of an MNLFA model and account for the data dependency in repeated measures (Greenbaum et al., 2013; Wang et al., 2013).

However, mixed evidence concerning the sufficient precision of a single calibration's results has led researchers to develop a multiple calibration (MC; Greenbaum et al., 2013; Henderson et al., 2013; Wang et al., 2013) approach. A MC method is used to reduce the uncertainty associated with a single calibration in a

conceptually similar manner to the popular multiple imputation (MI; Rubin, 1987; Schafer, 1997) method used to reduce the uncertainty associated with a single imputation. The MC technique was developed to produce more precise and stable estimators, with a precision and stability that increases as the number of calibrations increases. Though it is an intuitive extension of a single calibration technique, it also raises questions such as how to combine calibration estimates, and how many *m* calibrations may be needed to produce sufficient results (Greenbaum et al., 2013; Henderson et al., 2013; Wang et al., 2013).

It is an axiom in MI theory that the full information maximum likelihood (FIML) and MI estimates are equivalent when $m=\infty$ and same models are being tested (Graham et al., 2007). However, the necessary number of *m* needed to approximate $m=\infty$ remains unclear. Recently, recommendations set forth by Graham et al. (2007) suggested that much more than the previously believed 3-5 imputations are needed to produce sufficient results. By extending this question to encompass the MC method, a determination of the number of *m* calibrations are needed to produce a combined estimator that is sufficiently close to a maximum likelihood estimator (MLE) from the whole set needed to be established.

Wang et al. (2013) suggested to use 20 calibrations based on simulation results conducted on a longitudinal structured data set with one, fixed effect parameter. And in an ongoing IDA, Greenbaum et al. (2013) fit 20 latent growth curves to factor scores generated via calibration MNLFA models using data collected across 5 completed studies. .However, a large amount of inconsistency among calibration estimates was observed which may possibly be due to a larger number of model parameters being

tested, as well as additional sources of variation in real data that are absent in computer generated data. Additionally, the studies in the ongoing IDA have varying time points which can have confounding effects on latent growth curve parameters. Wang et al. (2013) motivated this study to extend previous simulation work to assess and compare the MC and MI estimators on longitudinal data with two, random effect parameters across various data characteristic combinations including sample size, effect size and number of time points per subject. Greenbaum et al. (2013) motivated this study to conduct a simpler version of the ongoing IDA for illustrative purposes.

Study Objectives

The objectives of this study were three-fold. First, this study aimed to compare MC combination rules to MI combination rules, and address the impact of data characteristics on the operating characteristics of these methods. The second objective was to determine if an additional parameter in the model and additional variation due to random effects would influence the previously recommended number of calibrations. The third objective was to demonstrate a successful implementation of an MC approach in an IDA framework with a simplified version of an ongoing project. The overall aim of this paper was to aid researches interested in conducting an MC technique by providing empirical evidence from an extensive simulation, an illustration from a real data application, as well as suggestions and considerations for future research.

This paper proceeds as the following: First, a brief literature review of traditional meta-analysis methods is given. Then, the novel capabilities of an IDA framework that offer the potential to powerfully extend pooled analysis techniques, but also rely on

calibration sampling, is depicted through portrayals of the model and data structures that are involved. Next, a simplified example of an MC technique is provided and a MC to MI comparison is made, which is then followed by a simulation that empirically evaluated the performance of these two methods in hypothetical scenarios using computer generated data. Finally, the MC approach is illustrated on real data in an IDA context data collected from three completed, randomized controlled trials that studied the effects of Multidimensional Family Therapy (MDFT; Liddle et al., 2002) to one of several active control treatment groups for adolescent substance use.

Chapter 2: Literature Review

Meta-analysis is a powerful method used to combine, analyze and statistically evaluate quantitative evidence from multiple independent studies to produce results based on a whole body of research (Hofer & Piccinin, 2009; Riley et al., 2007). A key characteristic that separates a meta-analysis from a literature review is its ability to examine the similarity of and potential reasons for dissimilarities of results across studies with quantitative rather than qualitative techniques (Blettner et al., 1999; Berlin et al., 2002). These techniques generally take two forms: aggregated data (AD) meta-analysis and individual participant data (IPD) meta-analysis.

Aggregated Data Meta-Analysis

Aggregated data (AD) meta-analysis refers to the statistical synthesis of *summary measures*. Summary measures are study- or arm-level values, or statistics, obtained from reports. In addition to the relative speed and inexpensive cost associated with performing AD meta-analyses, a main advantage is that summary data can be collected from a greater number of studies sought for inclusion than individual data can be collected from (Cooper & Patall, 2009). Although this method is ideal when raw data is inaccessible and may be sufficient for drawing some conclusions concerning the overall pattern of study-level characteristics affecting results, it is not suitable for conclusions concerning participant- level characteristics affecting results or any additional analyses testing new

research questions (Hofner & Piccinin, 2009; Nieri, 2003). Important cautions to be mindful of when conducting an AD meta-analysis including ecological fallacy (i.e., drawing inferences on individuals based on group results), and publication bias (i.e., the increased tendency (likelihood) that studies concluding positive results will be published while those concluding negative results are not) have been discussed in length, however, access to individual participant data has the potential to overcome them (Berlin et al., 2002; Cooper & Patall, 2009; Stewart & Tierney, 2002)

Individual Participant Data Meta-Analysis

Individual Participant Data (IPD) meta-analysis refers to the statistical synthesis of *individual comparable measures*. Individual comparable measures are subject-level values, or common items, measured either identically across studies or harmonized (altered to be made comparable) across studies (Hussong et al., 2013). Major advantages of this approach include a gain in power due to increased sample size, an ability to produce consistent analysis or test new hypotheses, and (iii) an opportunity to investigate the data directly and separate participant-level heterogeneity from study-level heterogeneity (Curran & Hussong, 2009; Fisher, 2011; Simmonds et al., 2005;). Although the process of IPD implementation – acquiring, checking and cleaning a large amount of data- may be difficult and very time- and labor-intensive, a vast amount of literature supports the relative benefits of IPD methods and regards IPD as the gold standard in systematic reviews (Bower et.al, 2003; Chalmers et al., 2002; Simmonds et al, 2005; Stewart & Tierney, 2002; Riley et al., 2007; Walveran, 2010).

Once the data has been obtained in an IDA, however, perhaps the most fundamental challenge is related to the consistency and quality of measures in each study. In order for a pooled analysis to take place, measures must be available from each study that reflect the same theoretical meaning and can be put on the same metric (Bauer & Hussong, 2009). Ideally, the same gold-standard assessment tool used to measure the outcome of interest would be used across all studies with alike validity and reliability. Realistically, clinicians are often confronted with the dilemma of choosing from a variety of assessment tools and the ability to reconcile the wide array of measurement practices used across studies is a common challenge in many areas psychological research (Curran &Hussong, 2009; Leccese & Waldron, 1994). Fortunately, replacing aggregated data with raw data permits the construction of complex data landscapes enabling sophisticated modeling techniques to perform a more flexible analysis (Glass, 2000).

In particular, a strategy borrowing from measurement invariance in factor analysis and linking and equating test scores in educational assessment has become of recent interest in clinical psychology to link studies together at the primary factor level; this strategy has been termed Integrative Data Analysis (IDA; Bauer & Hussong, 2009; Curran et al., 2008; Curran & Hussong, 2009; Hussong et al, 2013).

Chapter 3: Integrative Data Analysis

Integrative Data Analysis (IDA) refers to the statistical synthesis of *commensurate measures*. Commensurate measures are participant-level values, or scores, generated by the research synthesizer and constructed to have the same meaning and metric across studies, despite potentially significant between-study differences in modalities of assessment (Hussong et al, 2013). Using a latent variable approach, multiple indicators of the same construct measured across studies form a set of items used to simultaneously pool the data together through an assumed underlying factor; the existence of which is believed to have given rise to the pattern of correlations among the set of items (Bollen, 2002; Hussong et al., 2013). The latent variable, or factor, is unobserved and viewed as a common cause responsible for all of the observed item responses (Hussong et al., 2013). The relationship between each individual item with the factor is defined through a link function following a form (e.g., identity, logistic logarithm, etc.) determined by the distribution of that item (e.g., normal, binomial, count, etc.) that together form a system of equations. The distribution of the factor (e.g., normal) is set by the researcher so that analytic strategies can use a common underlying metric to calculate scale scores representing the construct of interest for all participants across all studies.

Ultimately, information from multiple measures is condensed into a single measure, rationalized by the assumption that the factor accounts for all the associations

among observed item responses. This technique, is sometimes referred to as local independence or data reduction, has widely been used in context of social and psychological research to assess abstract concepts of constructs that are inherently unable to be directly measured (Bollen, 2002). Thus, so long as there is a sufficient overlap of the item set across studies, IDA extends traditional IPD by allowing studies to have different indicators of a given construct and retaining indicators that cannot be harmonized across all studies, but nevertheless provide information within studies (Curran et al., 2013; Hussong et al., 2009

Table 3.1 Example Item Set Frequencies

	Items			
	Y1	Y2	Y3	Y4
Study 1	X		X	
Study 2		X	X	X
Study 3	X			X
Study 4	X	X	X	
Study 5	X	X		X

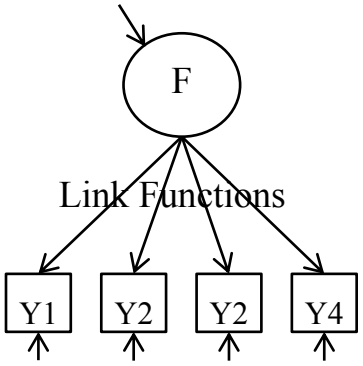


Figure 3.1 CFA Model

Table 3.1 and Figure 3.1 depict how studies can be linked together at the primary factor level using a uni-dimensional (one factor) Confirmatory Factor Analysis (CFA) model, in which the factor is set to be continuous and normally distributed. A large amount of resources regarding latent variable modeling is offered at website: www.statmodel.com (Muthén & Muthén, 2011).

Perhaps one of the most appealing capabilities of factor analytic frameworks, though, is the opportunity for the inclusion of a latent variable measurement model, such as MNLFA, that can condition properties of the model by observed predictors that would otherwise make a number of unrealistic assumptions of homogeneity in response distributions and probabilities.

Model Structure

A Moderated Nonlinear Factor Analysis model (MNLFA) extends traditional psychometric models by accommodating data with a variety of distributional properties (i.e., is generalized), and allows exogenous variables to moderate model parameters in three ways: to the factor mean, illustrated as blue lines; to the factor variance, illustrated as green lines; and to the relationship between the factor and an observed item, illustrated as red lines in Figure 3.2.

Significant covariate effects found in the factor mean and variance parameters specify conditional distribution indices for the model, while significant covariate effects found in the item parameters specify conditional probability indices for the model. Conditional distribution indices are sometimes referred to as impact and are accounted for by including regression terms in distribution specification functions that permit the

factor mean and variance parameters of the model to vary across observed predictors. Conditional probability indices are often referred to as differential item functioning (DIF), or factorial noninvariance. They are accounted for by including regression terms in the link functions of the model that permit the relationship between the factor and an item to vary across observed predictors (Bauer & Hussong, 2009). More specifically, mean parameters are expressed as a linear function of the moderators, variance parameters are expressed as a log-linear function of the moderators, and DIF parameters are expressed in the form of the specified link function. In a link function, covariates can act on the intercept, modifying the difficulty of the item, or on the slope (loading), modifying the discrimination of that item.

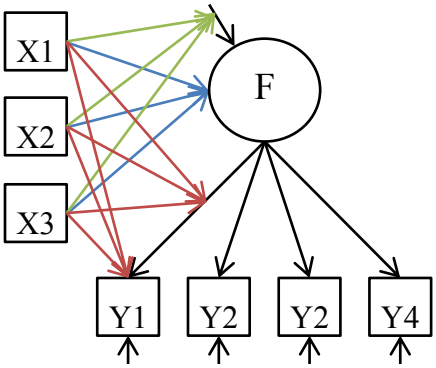


Figure 3.2 MNLFA Model

Once an MNLFA model has been established, commensurate measures in the form of factor scores are created for use in subsequent analysis. More specifically, maximum a posteriori (MAP) factor scores (i.e., the mode of the latent factor posterior distribution for each person j) are derived from the observed data through the model (Greenbaum et al., 2013). There is a comprehensive four step procedure including: preliminary feasibility analysis, selecting an item set, developing a measurement model,

and scoring available to guide researchers interested in conducting an IDA provided by Hussong et al. (2013) . Also, Bauer & Hussong (2009) provide a review of traditional psychometric models and an in-depth description of MNLFA model, as well as an IDA illustration, conducted on studies measuring alcohol involvement.

Given a brief insight to the complexity of the model structure, the following was focused on the complexity of the data structure.

Data Structure

The complications associated with an analysis on a pooled, multi-study data set stem from the automatic clustering of subjects by study, with the potential for further clustering within each study (e.g., students within schools; clients within clinics), which the individual-level characteristics of interest will be nested within

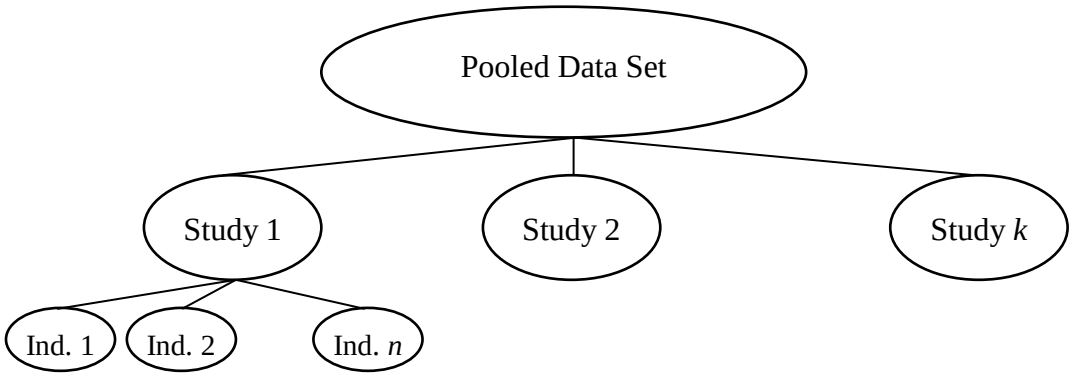


Figure 3.3 Pooled Data Structure

The hierarchal structure depicted in Figure 3.3 gives an idea of the nesting that correlational associations that must be accommodated when pooled data is analyzed. Moreover, an additional level of complexity is added when data is collected from

longitudinal studies, as within-subject correlations must also be taken into account. Barriers encountered when handling longitudinal data arise from the property that allows each subject to be measured repeatedly, of which the number and length between measurement occasions contributed by each subject to the whole data set can vary drastically both within and across studies, and exogenous variables can be either constant in time (e.g., gender, ethnicity) or changing in time (e.g., age, education level, marital status). This added dimension in the data configuration complicates correlational structures used in model estimation, and consequently, the computational power required to perform the necessary levels of integration in an MNLFA model is increased. It is a result of this intractability that led to a calibration approach that then led to the multiple calibration approach.

Chapter 4: Multiple Calibrations

The benefit of a method using more than one calibration stems from the ability to to use more information available in the whole set and capture the variability of one calibration to the next. Calibrations are constituted as random selections so it is unlikely that the same calibration will be drawn more than once. Thus drawing multiple calibrations would result in a set of subsets that uniquely represent the whole set, but that also are not completely independent from each other. Appropriately combining calibration estimates has the potential to produce an estimator that is close to MLE estimates produced by the whole set.

Simplified Example

Consider a longitudinal structured data set, R, with a sample size of N subjects for which each subject has up to T repeated measures of the outcome of interest, Y. Suppose we are interested in the trend of Y over time (slope), and that we also know the true trend, β, from the population, P, from which R was drawn. A simple linear regression model fit to R is not an appropriate analytic tool because repeated observations (level-1) over time within subjects (level-2) are flattened to a single level which fails to account for within-subject correlation (i.e., dependent observations are evaluated as independent). Ignoring the data dependency could result in misleading conclusions with biased parameter

estimates and degraded standard errors because multilevel data requires a multilevel approach (Kim et al., 2012). Let b_{ML} denote the estimate from an appropriate multi-level approach (i.e., mixed effects model maximum likelihood estimate) on the whole sample. The b_{ML} estimator is considered the gold standard for comparison because it is as close to the population parameter that the MC estimator can get. This concept was shown in Figure 4.1, which also depicted the concern encountered for how many calibrations will produce an estimate that is satisfactorily close to $b_{ML,}$ and by extension β.

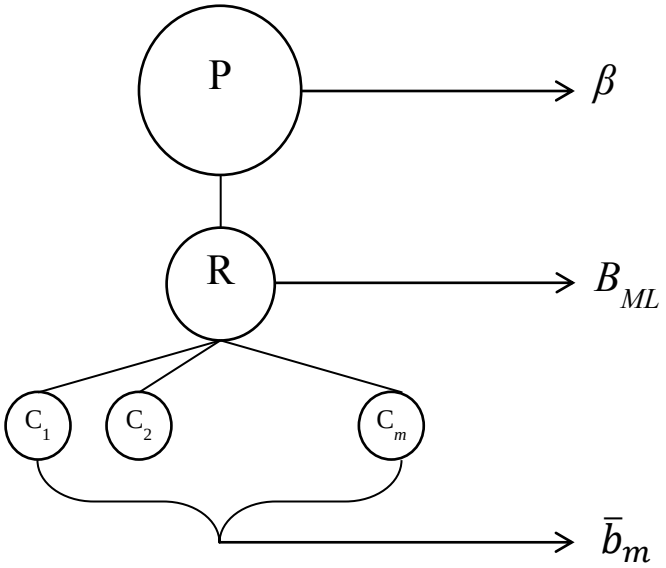


Figure 4.1 Multiple Calibrations

To calculate the multiple calibration estimate, first consider a single calibration, $C_{1,}$ drawn from R which will only contain a fraction of roughly $1/T^{th}$ (or more precisely the number of subjects/total number of measures) of the information available in *R*. However, because there is only one measure per subject, a simple linear regression model can be fit to C_1 without violating the key independence assumption. Denote this estimate

as b_1, and let us repeat this process m times and denote the corresponding estimates as b_1 to b_m. The significance of any single calibration estimate is determined through a model fit directly to the calibration. Calibration estimates are subject to chance variability due to random selection, and to a power loss due to a decrease in sample size from the whole set to the calibration subset. However, because the same model was fit across multiple calibrations, combining estimates obtained from each offers the potential to capture the variation among calibrations, and to substantially increase the power.

The estimator, $\bar{b}_m$ from multiple (m) calibrations represents the combined mean of all the calibration estimates, and is simply the calculated as the average of all b_k, $k \leq m$. The variance was partitioned into two parts: the within calibration variance,U_{wtn} which represents the usual type of sampling variability, and the between calibration variance,U_{btw} which captures the variability from one calibration to the next. These calculations can be expressed as:

$$\bar{b}_m = \frac{1}{m}\sum_{k=1}^{m} b_k \quad (1)$$

$$U_{wtn} = \frac{1}{m}\sum_{k=1}^{m} s_{b_k}^2 \quad (2)$$

$$U_{btw} = \frac{1}{m-1}\sum_{k=1}^{m}\left(b_k - \bar{b}_m\right)^2 \quad (3)$$

Analogous to the uncertainty associated with subsetting a longitudinal set is the uncertainty associated with filling in missing data of a longitudinal set.

Multiple Imputation

Multiple imputation (MI) is a method that combines estimates to produce final results for inference purposes in a way that is conceptually similar to the MC method.

While measuring participants repeatedly over time is a powerful method to estimate rates of change over time, it also provides repeated opportunities for participants to drop out or to miss measurement occasions (Siddique, 2012). Because many common statistical analyses are designed with complete data in mind, researchers have become increasingly aware of the problems and biases which can be caused by missing data (Merkle, 2011). Therefore, filling in, or imputing, missing values based on the observed data to generate a complete data set to use for subsequent analyses has become an attractive option (He & Raghunathan, 2012). However, using only one imputed dataset ignores the uncertainty involved with replacing missing data (Merkle, 2011). To formally address this uncertainty, a widely accepted and flexible approach is to fill in each missing datum with several (m) sets of plausible values, conduct separate but equivalent analyses on each of the completed datasets, and combine results to produce the final estimators used for inference (He & Raghunathan, 2012; Merkle, 2011; Siddique, 2012). The MI method is similar to the MC method only in that final results are obtained by combining a set of estimates, each of which are associated with some uncertainty in the belief that the combined estimate will be more precise.

Table 4.1 Multiple Calibration to Multiple Imputation Comparison

	Multiple	
	Calibration	Imputation
Problem	Longitudinal data structure	Missing data mechanism
Solution	Draw calibration and analyze a subset	Impute missing data and analyze a complete set
Reduce Uncertainty	Repeat m times Combine results for inference	Repeat m times Combine results for inference

Table 4.1 shows a comparison between MC and MI techniques. Though estimating the uncertainty in the MC case is inherently different from the MI case, the point estimate and within and between variance are computed equivalently and the combining aspect makes it the closest statistical method available for comparison. The formulas set forth by the MC and MI methods for the total variance and degrees of freedom of a combined mean estimator are given in the following section, along with a simulation addressing some of the most prominent concerns pertaining to power, precision, and solution stability.

Chapter 5: Simulation

Methods

This simulation evaluated and compared MC estimators to MI estimators in a cross classified design that also assessed the impact of effect size, sample size, and number repeated measures per subject, or time points. The point estimate and the within and between variance for the MC and MI methods were computed equivalently as expressed in equations (1-3). The total variance, U_{MC}, and degrees of freedom, df_{MC} , using MC combination rules developed by Wang et al. (2013) are:

$$U_{MC} = \max\left(U_{\text{wtn}} - \frac{T}{T-1}U_{btw}, \quad 0\right) + \left(\frac{T}{T-1} - \frac{m-1}{m}\right)U_{btw} \tag{4}$$

$$df_{MC} = \max((n-2)m - (n-1)/T, \quad nT-1) \tag{5}$$

Where *n* represents the number of subjects; *T*, the number of time points possible per subject; and *m*, the number of calibrations. Then, the formulas using the MI combination rules for the total variance, U_{MI}, and degrees of freedom, df_{MI}, adopted from Schafer (1997) are:

$$U_{MI} = U_{\text{wtn}} + \left(1 + \frac{1}{m}\right)U_{btw} \tag{6}$$

$$df_{MI} = (m-1)\left(1 + \frac{mU_{wtn}}{(m+1)U_{btw}}\right)^2 \tag{7}$$

Data Generation

A total of 27 2-parameter linear model scenarios where investigated, for which the parameter of interest being combined was the slope. The investigated model scenarios were 9 longitudinal structured data sets representing all combinations of *T*=4,8 and 12 repeated measures per subject with effect sizes of *D*=0.3, 0.6 and 0.9 were each evaluated using sample sizes of *N*=100, 200 and 500. The entire simulation was conducted in SAS (Version 9.3), and all seeds used in the random number generators are available Tables A.3 and A.4 of Appendix A. Each population set was created with N=100,000 subjects assuming population (fixed) effects $\beta = (\beta_{0,} \beta_1)^T$, random effects $b_i = (b_{i0},\ b_{i1})^T \sim N(0,\ \boldsymbol{\Sigma})$, subject specific effects $\beta_i = (\beta_{i0,} \beta i_1)^T$ and random error terms $\varepsilon_{ij} \sim N(0,\ \sigma_\varepsilon^2)$. The observed measure Y_{ij} for subject *i* at time *j* were expressed:

$$Y_{ij} = \beta_{i0} + \beta_{i1} t_j + \varepsilon_{ij},\ i = 1,\ 2, \ldots, N, j = 0,\ 1, \ldots,\ (T-1) \quad (8)$$

$$\begin{cases} \beta_{i0} = \beta_0 + b_{i0} \\ \beta_{i1} = \beta_1 + b_{i1} \end{cases}$$

All random intercept and error variance components were set to 1.00. The slope variation was set to be one-tenth of the intercept variation, and subject specific intercepts and slopes were correlated at .5 in attempt to mimic real data (i.e., $\sigma_{b_0}^2 = 1$, $\sigma_{b_1}^2 = 0.1^2$, $cov(b_0,\ b_1) = 0.05$ and $\sigma_\varepsilon^2 = 1$). The population intercept was set to zero in all scenarios. The population slopes were calculated by the difference between the first and last time points divided by standard error at baseline, using effect size formula,

$D = \frac{y(t_{T-1}) - y(t_0)}{\sqrt{\sigma^2_{y(t_0)}}}$, suggested by Feingold (2009) for growth modeling. After the population sets were generated, the parameters were recovered using SAS proc mixed. These estimates (Table A.1 of Appendix A) were very close to the values used to generate the sets (Table A.2 of Appendix A) and were considered to be the "true" slope parameters used in all future calculations involving β_1. The nine population sets are depicted graphically in Figure 5.1

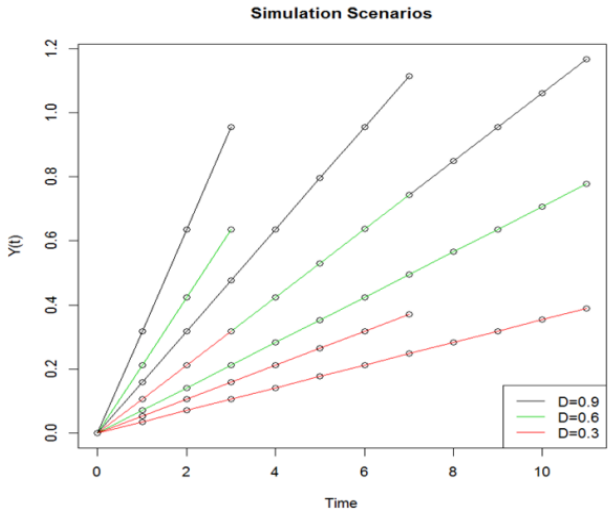


Figure 5.1 Simulation Scenarios

It is important to note that in this simulation, a change in time points was also associated with a change in time span, not in the density of the points. When calculating slope parameters, time and effect size were fixed, so the slope for a given effect size over four time points compared to the same effect size over eight time points would have a larger magnitude because it was reflecting the same amount of change in a shorter period of time. Though frequency of measurement occasions is often an important feature

considered when designing a study, in the context of an IDA this decision will not be up to the analyst becuase secondary data is being used.

Next, N subjects representing the sample size under consideration (N=100, 200, and 500) were selected from the corresponding population set using SAS proc surveyselect. Each replication set R_r , r = 1,2..500, was drawn without replacement from the population set, however the same subject could be drawn for two or more replications. The maximum likelihood estimate of the slope, $b_{MLE,}$ from a mixed effects (random intercept, random slope) model for each set was calculated using SAS proc mixed and considered as close to the population parameter as the calibration estimate could get. To compute the single calibration estimates, SAS proc surveyselect was used again to randomly select one observation per subject to make C_k, k=1,2,…100 independent subsets of R, and a simple linear regression model was fit to each calibration using SAS proc reg. The m^{th} calibration combined estimate was calculated in the order the calibrations were drawn. The MC formulas for the standard error and degrees of freedom of the combined estimate were calculated parallel to the MI formulas and used for inference according to equations 1-7.

Performance Measures

To conduct hypothesis tests with MC and MI estimators, the MC method followed a $\frac{\bar{b}_m}{\sqrt{U_{MC}}} \sim t(df_{MC})$ distribution, and the MI method followed a $\frac{\bar{b}_m}{\sqrt{U_{MI}}} \sim t(df_{MI})$ distribution. The measures used to assess the performance of this method included bias, mean square error (*MSE*), and Type II error rate for both the multiple calibration and the multiple imputation formulas calculated with the following formulas:

$$Bias = E\left|\hat{b} - \beta\right| \tag{9}$$

. $$MSE = \left(E\left(\hat{b}\right) - \beta\right)^2 + VAR(\hat{b}) \tag{10}$$

$$\text{Type II error} = \begin{cases} 1, & \frac{\hat{b}}{SE(\hat{b})} < t_{df}^{.975} \\ 0, & \frac{\hat{b}}{SE(\hat{b})} > t_{df}^{.975} \end{cases} \tag{11}$$

Type I error was set to .05, and power was calculated as the proportion of the 500 replications that correctly rejected the false null hypothesis (i.e., when $\frac{\hat{b}}{SE(\hat{b})} > t_{df}^{.975}$). and the relative efficiency was calculated as a ratio of *MSE*s.

Results

The simulation results were presented through the evaluation of performance measures in order of power, degrees of freedom, standard error, bias, mean square error (MSE) and relative efficiency.

Power

A single calibration method was severely underpowered. The commonly desired 80% power level was never reached for small effect sizes, and only reached for medium effect sizes with 500 subjects, at least 200 subjects were necessary for large effect sizes, irrespective of the number of time points over which it was observed. Table 5.1 summarizes the observed power and 95 % confidence intervals when implementing a single calibration method.

Table 5.1 Power and 95% Confidence Intervals of a Single Calibration

		T=4		*T*=8		*T*=12	
		p	(95% CI)	p	(95% CI)	p	(95% CI)
D=.3							
	N=100	0.1	(.07, .13)	0.14	(.11, .17)	0.1	(.07, .13)
	N=200	0.21	(.17, .25)	0.21	(.17, .25)	0.18	(.15, .21)
	N=500	0.46	(.42, .50)	0.4	(.36, .44)	0.37	(.33, .41)
D=.6							
	N=100	0.36	(.32, .40)	0.33	(.29, .37)	0.3	(.26, .34)
	N=200	0.67	(.63, .71)	0.54	(.50, .58)	0.52	(.48, .56)
	N=500	0.96	(.94, .98)	0.93	(.91, .95)	0.87	(.84, .90)
D=.9							
	N=100	0.61	(.57, .65)	0.6	(.56, .64)	0.54	(.50, .58)
	N=200	0.94	(.92, .96)	0.89	(.86, .92)	0.84	(.81, .87)
	N=500	1		1		0.99	(.98, 1)

Multiple Calibration Combination Rules

Implementing a multiple calibration method, the combined mean estimate from three calibrations was sufficient to achieve 80% power when the effect size was large, regardless of time span or number of subjects. For a medium effect size, 80% power was reached in 3, 5 and 15 calibrations for 500, 200 and 100 subjects, respectively. Lastly, for small effect sizes, 10 calibrations were sufficient for 500 subjects, about 65 calibrations were necessary for 200 subjects with 8 or 12 time points to reach 80% power. However, combined estimate from 100 calibrations never achieved 80% power in the case of a small effect size with 4 time points and 200 subjects, or for any number of time points with 100 subjects. More specifically, $\bar{b}_{100}$ had about 60% power for 4 time points and 200 subjects, and about 20%, 45% and 50% power for 4, 8 and 12 time points and100 subjects, respectively.

Multiple Imputation Combination Rules

Implementing a multiple calibration method and using the multiple imputation rules to estimate the combined mean standard error and degrees of freedom, power was observed as expected only in scenarios with a medium effect size and five hundred subjects, or with a large effect size and at least two hundred subjects. The observed power for these scenarios was consistently lower than the power calculated using the multiple calibration formulas.

In all other scenarios (small effect sizes and medium effect sizes with 200 subjects or less), there was an interesting occurrence in which power appeared to decrease as the number of calibrations increased. The pattern was exaggerated at higher effect sizes and can be explained by situations in which the first few calibration estimates were close together yet far from the null-hypothesis, which results in a large point estimate and a small between calibration variance. Recall the degrees of freedom formula in equation (7), small U_{btw} and m yielded an extraordinarily high degrees of freedom that when combined with a large point estimate the null-hypothesis is rejected causing power to be artificially high before dropping and leveling off. This phenomenon was also observed in the simulation study by Graham et. al. (2007) assessing suitable number of imputations in multiple imputation theory.

Figure 5.2 showed the increase in power associated with increase in number of calibrations when using the MC combining rules, as well as the observed "power drop-off" when using the MI combing rules. Tables A.5 – Table A.7 in Appendix A provide the average observed degrees of freedom for selected calibrations in small, medium and large effect size scenarios, respectively.

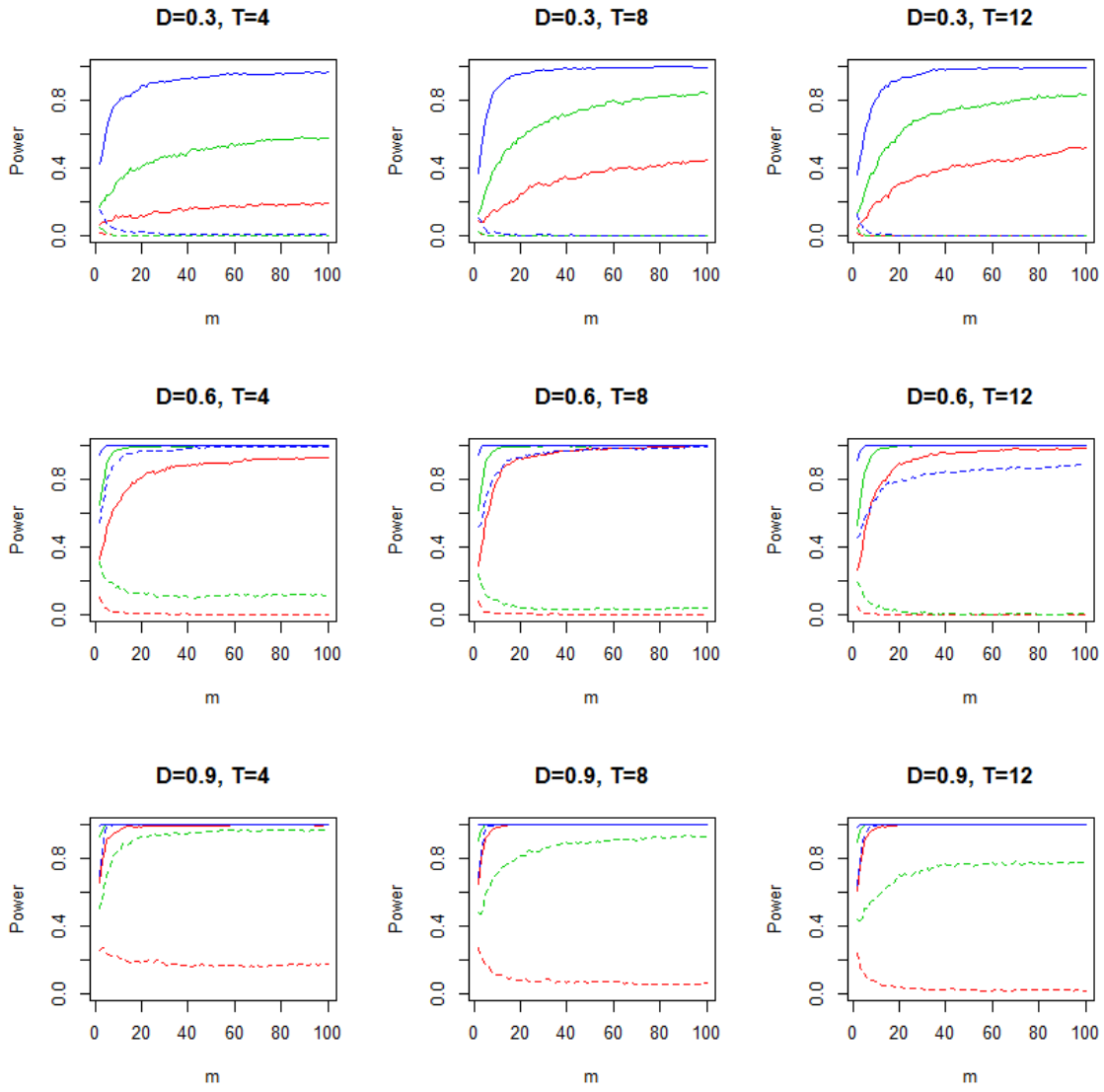


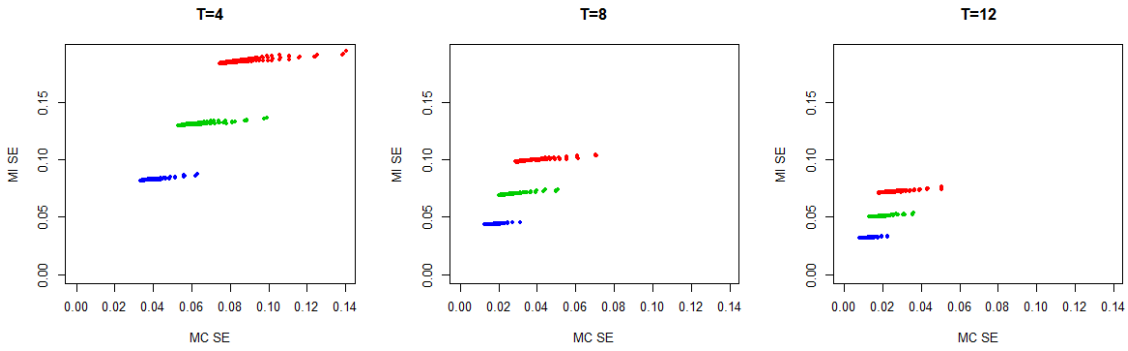
Figure 5.2 Power Analysis

The instability of the degrees of freedom at lower levels of *m* using the MI formula was clearly shown in Tables A.5 to Table A.7 of Appendix A, whereas the degrees of freedom when using the MC formula steadily increased as *m* increased, and in most cases approximates the normal distribution. Due to the role of degrees of freedom effect on power, using the *t*-statistic perhaps provides a clearer picture of evaluating power. Figure A.1 in Appendix A shows that as the number of calibrations increases, the

t-statistic rises as expected using the multiple calibration formulas whereas it appears to be relatively constant using the multiple imputation formulas.

Standard Error

The relationship between the MC and MI standard errors are shown in Figure 5.3. Each cluster of points in Figure 5.3 consisted of all 3 effect sizes, as effect size had no influence on any other performance measure except power. Then, regardless of the sample size, the estimated standard error using the MI formulas was consistently about 2.3, 3.0 and 3.4 times the estimated standard error using the MC formulas in all 4, 8 and 12 time point scenarios, respectively.



N=100 *N*=200 *N*=500

Figure 5.3 MC vs. MI Standard Errors

The standard error using both the MC and MI formulas were reduced by about 30% and 37% when the sample size increased from 100 to 200 and from 200 to 500, respectively. However, the observed reductions in the two standard error estimators were not the same when the numbers of time points were increased. The MC estimates were reduced by about 58% and 35%, while the MI estimates were only reduced by about 46% and 27% when the number of time points increased from 4 to 8, and 8 to 12, respectively.

Bias

The absolute bias at any level of *m* was systematically reduced by about 30% when the number of subjects was increased from 100 to 200 and when the number of subjects was increased to 500, absolute bias was reduced by an additional 30%, irrespective of the number of time points. Then, similarly, the absolute bias was reduced by about 50% when the number of time points increased from 4 to 8 and when the number of time points was increased to 12, absolute bias was reduced by an additional 10%, irrespective of the number of subjects. The asymptotic tendency for the bias associated with the multiple calibration estimator to approach the bias associated with the MLE estimator as *m* approaches 100 is illustrated in Figure 5.4

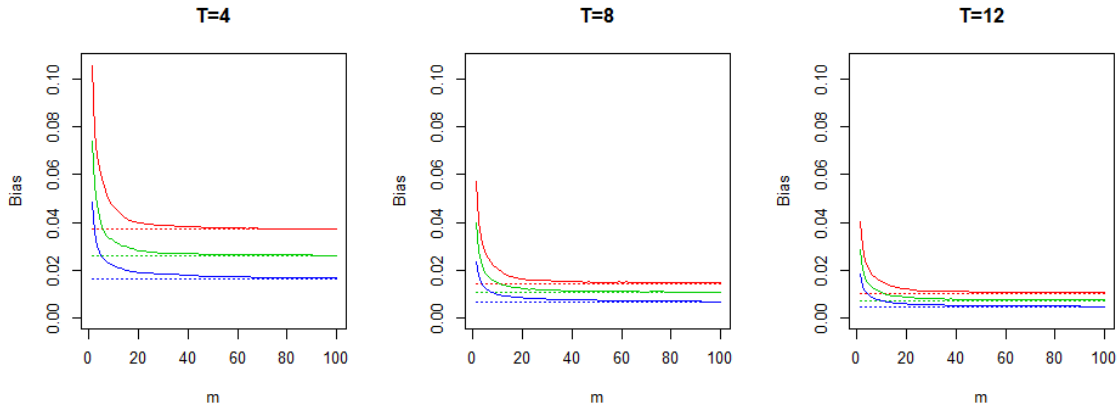


Figure 5.4 Bias Analysis

An initial steep drop and an apparent leveling off effect after around 20 calibrations was depicted in all scenarios. This suggested that even small numbers of *m* can significantly reduce the bias (and hence increase the precision) of the calibration

estimate, and that the relative contribution of an additional calibration in doing so decreases as *m* increases, and further starts to diminish around *m* = 20.

The rate at which the percentage of additional bias associated with a single calibration estimate was reduced when implementing a multiple calibration approach was consistent across sample size, effect size, and number of time points. Table 5.2 and Figure 5.5 represents the percentage of additional bias from MLE estimator to single calibration estimator that was reduced by using a multiple calibration estimator instead.

Table 5.2 Percentage of Single Bias Calibraiton Reduced

Number of Calibrations	% Bias Reduction
100	95
80	90
40	90
20	80
10	70
5	60
3	45
1	0

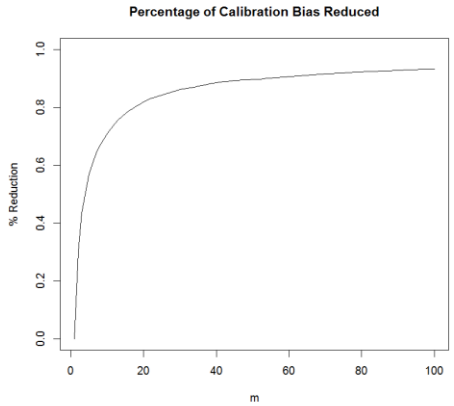


Figure 5.5 Percentage of Single Calibration Bias Reduced

When compared to bias of a single calibration estimate, a combined, 5 calibration estimator reduced the bias by more than half, and by 100 calibrations nearly all of the added bias was removed. The value of an additional calibration to the combined estimator was much higher at smaller values of *m*. For example, the second and third calibrations together reduced the bias of the estimator at m=1 by about a half, while the fourth and fifth calibrations together only reduced the bias of the estimator at m=3 by about a sixth. Or alternatively, 20 calibrations reduced the added bias by about 80% while 40 calibrations (an additional twenty) reduced the added bias by about 90% (an additional 10%). So even though the first twenty and second twenty calibrations reflect the same amount of work, there is much less reward in the latter.

Mean Square Error

Mean Square Error (*MSE*) is a measure that incorporates both the bias and standard error of the estimator. Figure 5.6 shows the plots of *MSE*s for the MC and MI estimators. The asymptotic tendencies where very similar to those observed in the bias plots, again showing that as *m* increases, the contribution of an additional calibration to the combined estimator decreases. About half of the added precision was seen within the first 5 calibrations, and the contribution of calibrations when m>20 was minimal and diminished as *m* approached 100. The *MSE* of the MI estimator leveled off at higher values than the MSE of the MC estimator in a pattern consistent with relationship observed between the MI and MC standard errors shown in Figure 5.3

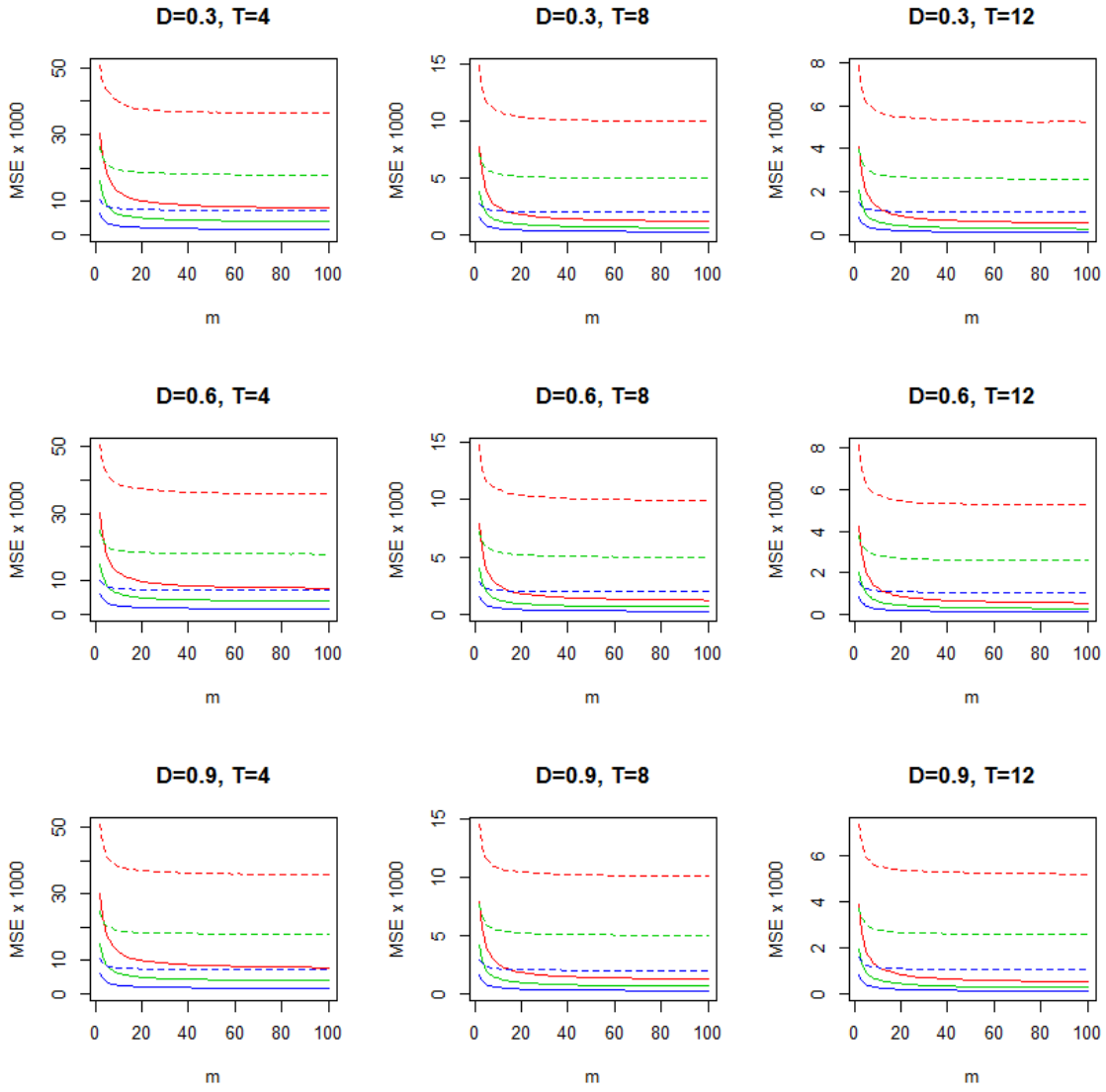


Figure 5.6 ***MSE* Analysis**

Relative Efficiency

The relative efficiency of the multiple calibration estimator was calculated as a ratio of *MSE*s and was considered in relation to both the random and fixed effects MLE estimators, as well as to the multiple calibration estimator at m=100.

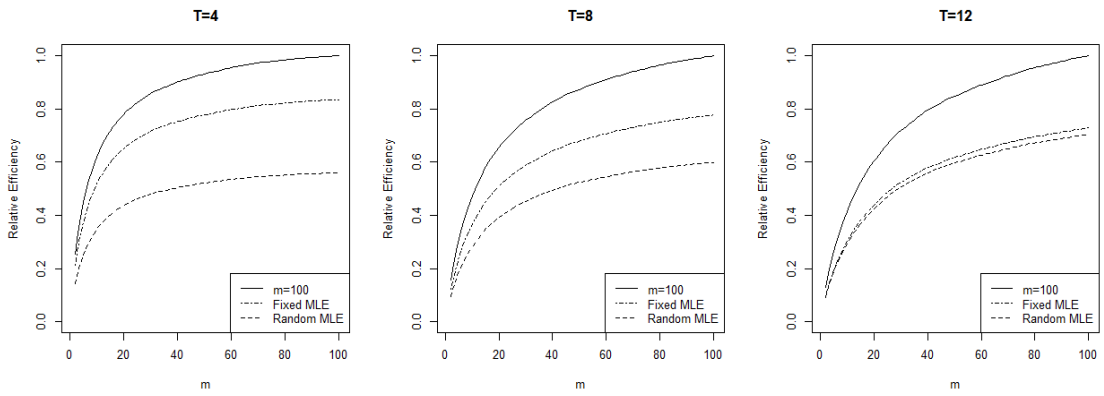


Figure 5.7 Relative Efficiency

For smaller levels of *T*, the MC estimator was not as efficient as the MLE estimators, and the efficiency in relation to the 100th calibration had a more rapid initial increase and quicker leveling off tendency than when compared to the efficiency of the MC estimator at higher levels of *T*. Also, the higher relative efficiency of the fixed effects MLE indicated the multiple calibration method is more suitable for data assuming fixed rather than random effects. However, this difference diminished as *T* increased. The MC estimate is calculated in the same way whether the whole data set is assumed to have fixed or random effects. The estimate just won't be as efficient for a fixed effects data set as it is for a random effects data when the number of time points is small.

Summary

Simulation findings demonstrated that a combined mean method using the MI combining rules are inappropriate to use in the case of a multiple calibrations. There were strong asymptotic tendencies observed for the MC estimates to approach the MLE estimates as *m* approaches 100 with respect to bias and *MSE*. The relative value of a calibration was much higher for smaller levels of *m*, as about half of the benefit gained

when implementing a multiple over single calibration was seen within 5 calibrations. and then the appeared to level off after about 20 calibrations.

Of the data characteristics investigated, an increase in effect size reflected an increase in power and had no effect on any other performance measure. An increase in sample size increased the power, reduced the bias, systematically reduced the MC and MI standard error estimates by the same amount, and had no effect on relative efficiency to the *MSE* estimator. An increase in time points increased the power, reduced the bias, reduced the MC and MI standard error, allowing for a greater reduction in MC estimates than in MI estimates, and increased the relative efficiency with respect to the MLE estimator.

The suitable number of calibrations will depend on the data at hand and goal of the analysis. If the main objective is to detect significance of a trend over time, strong effect sizes and large sample sizes will pick up the significance fairly quickly and five calibrations will be sufficient in many scenarios. Five calibrations was also observed to reduce the additional bias and inflated standard errors of a single calibration by about half, while 20 calibrations while stability in terms of how much an additional calibration would impact the combined estimate was observed at around 20 calibrations. However, if the main objective is to determine a precise estimate, one hundred calibrations would reduce all but 5% of the additional bias associated with a calibration approach compared to the MLE approach.

Limitations and Future Research

All data was generated with random effects, meaning that each subject had a unique intercept and slope, and the parameters were normally distributed. If the data was generated so that each subject had same intercept and slope, and the parameters were fixed, it is likely that less calibrations would be needed as random effects would be removed and the variation in the data would reflect only measurement error. However, the extent to which the within and between subject variation affects the multiple calibration method was not examined because variance parameters were held constant for each simulation scenario.

In this simulation study, time was measured discretely as integers ranging from 0 to 11. It is likely that continuously measured time would add accuracy in the regression model fit to calibration samples and yield better results. Also an increase in number of time points reflected an increase in time span, not in the number of points within a given interval. Therefore, it is in the same way likely that an increased point density would add accuracy to the regression model and yield better results. Rather than changing the effect size and number of subjects, which only change the scaling of performance measures and associated power, future investigation of the effect of point density and variance on the number of calibrations needed using simulation studies is recommended.

Lastly, the simulation analysis only used complete data sets in which every subject contributed the same number of measurements to the whole set to reduce likelihood of using and interpreting biased results. In reality, a missing data mechanism will likely be present. If subjects with fewer recorded measurements are systematically different than the subjects with more recorded measurements, the combined calibraiton

estimate may be biased towards the measures of those with fewer points. For example, if a particular subject only contributed one measurement occasion to the data set, that one measurement would be selected for every calibration, and the resulting probability of selection into a calibration sample for that measurement is 1. Whereas, for a subject with T measurements ($T>1$), the probability of selection into a calibration sample for those measurements is $1/T$. Missing data is well known to cause a multitude of problems, so the possibility for missingness to render misleading results when implementing a multiple calibration method must be recognized. Multiple imputation, the method from which combining rules were borrowed for comparison to the multiple calibration combining rules was indeed developed to handle uncertainty associated with missing data. However, as noted previously, MI is inherently different than MC, and the MI combining rules perform very poorly in an MC scenario. Implementation and performance of the multiple calibration approach combined with multiple imputation approach has not yet been examined.

The current project is somewhat different than the simulation, as the simulation is an extremely simplified example of how a calibration is used to estimate an MNLFA model in an IDA. In the case of multiple calibration approach in an IDA framework, rather than combining estimates fit on an actual calibration itself, there is an intermediate step involved. The calibration estimates being combined are actually fit to a longitudinal set, but the longitudinal set is constituted as factor scores generated using an MNLFA model that a calibration estimated. So, though information from the whole set is being used in the model for the estimates being combined, each set of scores still reflects a calibration's unique representation of the whole set.

In the MDFT application, a two parameter linear latent growth curve was fit to four discrete time points on a normally distributed outcome representing substance use. Implementing a multiple calibration approach in this case, the main objective was to determine the significance of parameter estimates, rather than the actual values of the parameter estimates, because there was no explicit clinical interpretation of meaning behind a substance use factor score. Also, due to the collapsed nature of factor scores, stronger effect sizes would provide a stronger indication that the observed treatment effect wass not just statistically significant, but makes a meaningful difference in reducing substance use for adolescents in a real world setting as well. Regarding sample size, there were 401 subjects in the whole sample, with subgroup sizes ranging from 57 to 311. For these purposes and based on the simulation results, five calibrations should be sufficient.

Chapter 6: MDFT Application

Background

Substance Use

Substance use among early adolescents is a significant public health concern as it is among one of the most robust predictors of severe substance use, criminality, and pervasive difficulties across life domains in later adolescence and adulthood (Burleson & Kaminer, 2007, Liddle et. al, 2009). While research for adolescent drug problems has increased, it is still sparse when compared with research of other adolescent problems, such as anxiety, ADHD, and depression (Liddle, 2008). Although recent reviews have indicated that some evidence based interventions are effective across ethnicity and gender subgroups, the question of whether they are equally beneficial remains unclear because most studies lack adequate statistical power to detect either ethnicity or gender by treatment interaction effects (Henderson et al., 2013). Fortunately, the ability to combine and reuse data from completed trials has the potential to substantially increase both the sample size of these subgroups and subsequent power of the analysis to answer these questions by reusing data. However, as is a common challenge in many areas of psychological research, there is no blood substance use content and the lack of standardization in assessment tools results in disparate measures used across studies (Curran & Hussong, 2009; Leccese & Waldron, 1994). Nonetheless, this seeming limitation becomes a distinct advantage in an IDA framework because multiple

measurement methods used across studies can be incorporated to actually strengthen the assessment of underlying construct (Curran & Hussong, 2009).

To illustrate this recent progress in pooled data analysis methodology as well as to assess the effectiveness and to improve the understanding of potential interactive influences on substance use interventions, an IDA implementing a multiple calibration MNLFA approach was conducted on three randomized controlled trials that compared Multidimensional Family Therapy (MDFT; Liddle, 2002) to one of several active comparison treatments in minority youth.

Multidimensional Family Therapy

Multidimensional family therapy (MDFT) is a family-based intervention for adolescent substance use that targets multiple realms of a teens functioning and social environment with a comprehensive focus that can be clustered into four important domains: Adolescent domain, helps teens to engage in treatment, develop coping, emotion regulation and problem solving skills; Parent domain, engages parent in therapy and increases their behavioral and emotional involvement with the adolescent; Interactional domain, focuses upon decreasing family conflict and increasing communication; and Extrafamilial domain, fosters family competency within all social systems in which the teen participates (Liddle et al., 2002; Liddle et al., 2008; Liddle et al., 2010). The integrated approach of MDFT is theoretically, clinically, and operationally different from the comparison treatments (i.e., individual cognitive behavioral therapy, peer group treatment, and residential treatment) and is expected to have treatment effects with a longer durability. The remainder of the paper simply refers to the control treatments as TAU (i.e., Treatment As Usual).

Data

Data for this study was obtained from an ongoing IDA (Greenbaum et al., 2013) with the inclusion criteria that a study had all four of the following measurement occasions: baseline (treatment initiation), 4 months (treatment termination), 6 months (short term follow- up) and 12 months (long term follow-up). Any study that did not have all four time points was not included and any additional time points available in the included studies were not used for simplicity purposes. The three included studies along with abbreviations used in the context of this paper are: ART, (Liddle & Dakof, 2002) and ATM (Liddle et al., 2009) and TEM (Lidddle et al., 2008).

Demographics

There was significant age, $F(2,398)$=60.65 ,p-value<.001),and ethnicity, $\chi(4)$=138.08, p-value < 0.001) by study differences observed in the pooled data set. With regards to age, the ATM study had an average age that was about two years younger than the other two studies. With regards to ethnicity, there were a larger proportion of African Americans in the TEM study, and of Hispanics in the ART study. Age was not included in any of the analysis, and although the trajectories of substance use over one year follow-up were regressed on both study and ethnicity main and treatment interaction effects, confounding of study by ethnicity cannot be completely ruled out. The aggregated data set, however, showed no significant subgroup differences between the MDFT and TAU groups, and the sample of sizes of the ethnicity and gender subgroups were substantially increased from those in the single studies. Tables 6.1 and 6.2 show break down of demographics by study and by treatment.

Table 6.1 Demographics by Study

	Study					
	TEM		ART		ATM	
Categorical	*N*	(%)	*N*	(%)	*N*	(%)
Treatment						
MDFT	112	(50)	53	(50)	34	(49)
TAU	112	(50)	54	(50)	36	(51)
Gender						
Male	182	(81)	80	(75)	49	(70)
Female	42	(19)	27	(25)	21	(30)
Ethnicity						
African Am.	161	(72)	17	(16)	32	(46)
Hispanic	23	(10)	76	(71)	35	(50)
White	40	(18)	14	(13)	3	(4)
Continuous	Mean	(Std.)	Mean	(Std.)	Mean	(Std.)
Age	15.40	(1.23)	15.36	(1.09)	13.67	(1.18)

Table 6.2 Demographics by Treatment

	Treatment					
	MDFT		TAU		Total	
Categorical	*N*	(%)	*N*	(%)	*N*	(%)
Study						
TEM	112	(56)	112	(55)	224	(56)
ART	53	(27)	54	(27)	107	(27)
ATM	34	(17)	36	(18)	70	(17)
Gender						
Male	156	(78)	155	(77)	311	(78)
Female	43	(22)	47	(23)	90	(22)
Ethnicity						
African Am.	103	(52)	107	(53)	210	(52)
Hispanic	67	(34)	67	(33)	134	(33)
White	29	(15)	28	(14)	57	(14)
Continuous	Mean	(Std.)	Mean	(Std.)	Mean	(Std.)
Age	15.02	(1.36)	15.16	(1.34)	15.09	(1.35)

Outcome Measures

The primary outcome of interest was the underlying construct of substance use, that was measured across studies using a 4 indicator item set in a uni-dimensional confirmatory factor analysis (CFA) framework. Table 6.3 provides the indicator descriptions along with the specified CFA link functions.

Table 6.3 Indicators of Substance Use

	Indicator			
	TLFB	AXI	PEI	USS
Name	30 Day Time Line Follow Back	Problem Oriented Screening Instrument	Personal Experience Inventory	Urine Analysis
Description	Number of drug use in the last 30 days for 5 different substances	Number of substance use problem, tolerance, and withdrawal symptoms	29 Items measuring substance use problem severity and frequency	Presence of five different substances in urine
Scoring	Count from 0 to 150 (Higher counts indicate more use)	Count from 0 to 17 (Higher counts indicate more use)	Sum of 29 4-point items (Higher numbers indicate more involvement)	0 = Urine test negative 1 = Urine test positive
Conditional Distribution	Negative binomial	Negative binomial	Censored normal (Censored from below at 0)	Bernoulli
CFA Link Function	Logarithm	Logarithm	Identity	Logistic

The four indicators of substance use were: 30 day TimeLine Follow Back (TLFB; Sobell & Sobell, 1992), Problem Oriented Screening Instrument for Teens (AXI; Rahdert, 1991), Personal Experience Inventory (PEI; Winters & Henly, 1989), and a urine analysis for five substances (USS; benzodiazepines, cocaine, methamphetamine, morphine, and THC). The TLFB was a count variable scored from 0 to 150 based on the number of days that a participant had used any of the 5 drugs during the 30 day period. The AXI was a count and PEI was a continuous variable that were both scored according to typical procedures discussed in the test manuals. The urine analysis was a binary variable scored as 1 if any of the five substances was recorded as a positive test and 0 if all substances were negative. Histograms of each indicator at each of the four time points are shown in Figures B.5 to Figure B.8 of Appendix B. The TLFB and AXI items were analyzed assuming negative binomial distributions, the PEI item was analyzed assuming a censored normal distribution, and the USS item was analyzed assuming a binomial distribution. The spaghetti plots in Figure 6.1 traced the indications of substance use over time by subject for the TLFB, AXI and PEI items, and the item frequencies by study over time are in Table B.1 of Appendix B.

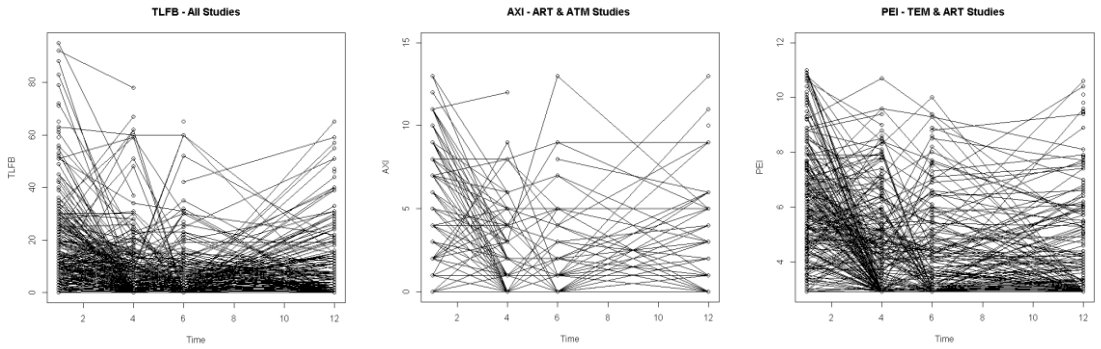


Figure 6.1 Indicator Spaghetti Plots

The upward trend between 6 and 12 months was not surprising as treatment effects measured close to treatment termination tend to be exaggerated, and it was estimated that about 60% of adolescents relapse within 3-12 months of completing a substance use intervention (Burleson & Kaminer, 2007). While longer follow-up periods would be critical in determining the sustainability of treatment effects, if MDFT altered the trajectories of adolescent substance use for at least 12 months there was cause for optimism (Liddle et al., 2002).

Means of all four indicators over overall, by study and by treatment are shown in Figure B.1 to B.4 of Appendix B. However, mean over time figures have the potential to mislead results when assessing longitudinal data. It is also noted that all of the included studies compared MDFT to active treatment control groups. Significant substance use reductions from baseline to 12 months were observed in all of the treatment arms, in all studies. Therefore, the fundamental question of this analysis was to determine if MDFT decreased substance use to a significantly greater extent than the TAU group.

Analytic Procedure

A 5 calibration MNLFA approach was conducted in an IDA to assess the effects of substance use interventions. First, five calibrations consisting of a single measurement occasion per subject were drawn and an MNLFA model was fit to each. Due to the complexity of an MNLFA model and the number of covariate effects that were tested, a series of 6 smaller models were tested first to reduce computational load and increase the chances of convergence. Figure 6.2 showed the all the covariate effects that were tested in the MNLFA models.

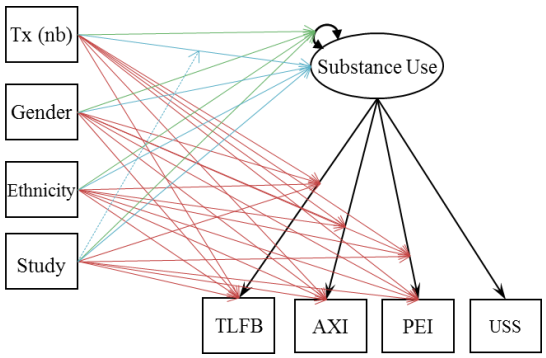


Figure 6.2 Substance Use MNLFA Model

The six smaller models were run in Mplus (Version 7) to trim and obtain starting values for the covariates effect estimates in a process similar to a backward-stepping selection procedure.

The first Mplus model run allowed the latent mean (blue lines in Figure 6.2) to vary by observed predictors and tested for potential latent variance moderators (green lines in Figure 6.2). Variance parameters not significant at the $p<.25$ level were dropped and the model was estimated again for the second mplus run. Only covariates that remained significant at the $p<.10$ level were retained in the variance component of the MNLFA model for this calibration, and estimates obtained here were later utilized as starting values for the final model.

In the next step of analysis, measurement invariance across subgroups for each of the self-report indicators of substance use was tested for potential differential item functioning (DIF; red lines in Figure 6.2). DIF was not tested for the urine analysis measure because of the objective biological nature of the measure presumably not subject to conditional probabilities relating to the factor. To test for DIF, the latent mean was allowed to vary by all observed predictors. The latent variance was fixed at one, and

covariate effects were tested on the intercept (difficulty) and slope (discrimination) of the link function for the TLFB, AXI and PEI items for the third, fourth and fifth mplus runs, respectively.

All DIF parameters found significant at the p<.05 level from these results were then tested together for the sixth and final mplus run. Estimates for the latent mean and DIF parameters from this model along with the variance parameter estimates found earlier were then used together as starting values in SAS (Version 9.3) nlmixed procedure to estimate the final calibration MNLFA model. Commensurate measures representing substance use in the form of maximum *a posteriori* (MAP) factor scores were then generated for the full set of observations by creating a 'dummy' indicator in the item set and using the SAS nlmixed procedure again. Further details for how nlmixed procedure was used can be found in Supplementary Material for Bauer & Hussong (2009).

This process was repeated 5 times, and equivalent latent growth curves (LGCs) were fit to each of the 5 sets of factor scores. A two parameter linear model was determined to be the best functional form where intercepts were regressed on study, gender, and ethnicity main effects, and slopes were regressed on all main effects, as well as treatment by study, treatment by gender, and treatment by ethnicity interaction effects. The structure of the LGC was shown in Figure 6.3 and the functions assessing substance use at time, *t*, with standard normal errors, ε, and dummy variables x_1 for treatment, x_2 for gender, x_3 and x_4 for ethnicity, and x_5 and x_6 for study were expressed:

$$F(t) = \beta_0 + \beta_1 t + \varepsilon \tag{12}$$

$$\begin{cases} \beta_0 = \beta_{00} + \sum_{k=2}^{6} \beta_{0k} x_k \\ \beta_1 = \beta_{10} + \sum_{k=1}^{6} \beta_{1k} x_k + \sum_{l=1}^{5} \beta_{1(6+l)} x_1 x_l \end{cases}$$

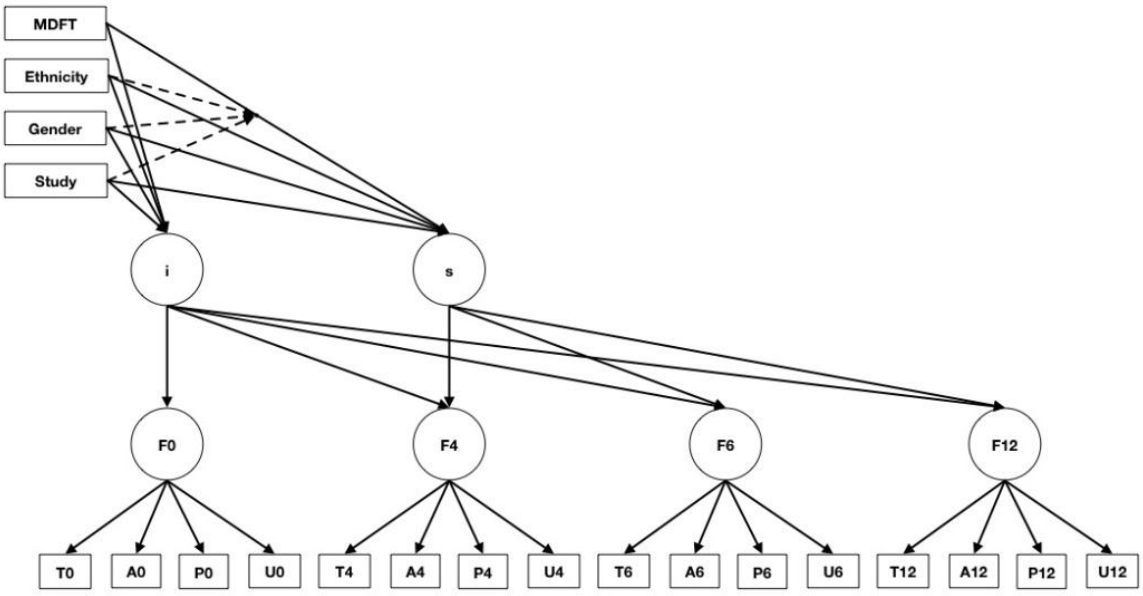


Figure 6.3 Substance Use LGC Model

Of particular interest were the MDFT-covariate interactions on the trajectories of substance use. The interaction effects were illustrated as dashed lines in Figure 6.3, and calculated using coefficients obtained in the second summation of the β_1 formula for equation (12). Point estimates and standard errors of the five LGCs were combined according to the MC combination rules to produce the final results

Results

The MDFT applications results are presented first through an examination of the calibration estimated MNLFA models and then though an evaluation of the individual and combined mean calibration LGC estimates

Calibration MNLFA Models

The observed differences in the calibration results were attributed to the variation due to chance in the random selection process of drawing a calibration. Table 6.4 showed

the frequencies of selected time points in each calibration and verified that each calibration reflected a unique representation of the whole set.

Table 6.4 Frequencies of Selected Time Points

	Time				
	Baseline	4 mo.	6 mo.	12 mo.	Total
Whole Set	400	306	301	317	1324
Calibration					
1	171	72	72	86	401
2	158	77	73	93	401
3	148	85	71	97	401
4	154	85	76	86	401
5	134	91	81	95	401
Total	1165	716	674	774	3329

To better understand what constituted the differences from one calibration to the next, as well as gain insight into the psychometric properties of the item set measuring substance, trends in the covariate effects found significant at the .05 level in the mean and variance impact and DIF parameters for the five MNLFA models were evaluated.

Mean Impact

Gender and ethnicity did not show significant impacts to the factor mean for any calibration. The treatment (non-baseline) impacts were always significant and negative direction, with MDFT decreasing the mean to a larger extant than TAU. The ART study was set as the reference group and impact of ATM was always negative and TEM always positive, which suggested that the least severe to most severe substance users by study went in the order: ATM, ART, and TEM. This was in agreement with the indicator mean graphs in Appendix B depicted.

Variance Impact

Treatment (non-baseline) was the only covariate effect included in all five of the MNLFA variance parameters, always in the positive direction suggesting that variation in substance use involvement increased after treatment initiation. The TAU non-baseline coefficient was always larger in magnitude than MDFT non-baseline. A likely reason for this was that the TAU group consisted of three different interventions combined, which consequently resulted in a wider range of effects on substance use trajectories. Ethnicity was indicated as significant variance moderator in one of the calibrations initial mplus runs (excluding DIF), however, when the final model was fit (including DIF) there no longer showed significance.

Differential Item Functioning

Evaluating DIF was more complicated than evaluating the mean and variance impacts because there were more parameters to consider and the interpretation of conditional probabilities was less intuitive than the interpretation of conditional distributions. Table 6.5 shows the frequencies of moderation effects included in DIF models.

Table 6.5 Frequency of DIF Included in the MNLFA Model

	Indicator					
	TLFB		AXI		PEI	
	Int.	Load.	Int.	Load.	Int.	Load.
Study	4	4	3	5	0	0
Tx (non-baseline)	4	5	4	5	1	1
Gender	1	1	1	1	1	3
Ethnicity	0	3	4	3	1	1

PEI appeared to function the most equivalently across subgroups, with the only observed DIF trend in the loading by gender which suggested that the relationship between the factor and PEI was stronger (higher/more discriminating) for females than for males. The MDFT and TAU treatment (non-baseline) effect estimates for both TLFB and AXI were all negative on the intercept and positive on the slope, suggesting these indicators became less difficult and more discriminating after initial baseline assessment. The study and ethnicity effects changed directions calibration to calibration.

Latent Growth Curves

The means of factor scores across each of the four time points to which the LGCs were fit were similar across the five calibrations. Figure 6.3 showed the mean trends over the one year follow-up for each set of factor scores. Table 6.6 provided MDFT point estimates and standard errors of the differences in treatment effects overall and within each gender and ethnicity subgroups for the LGCs individually, as well as combined. Figure 6.5 showed the estimated overall (main) MDFT and TAU effects.

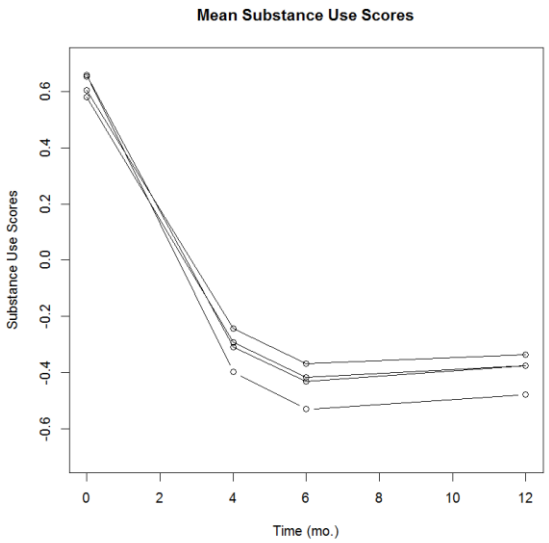
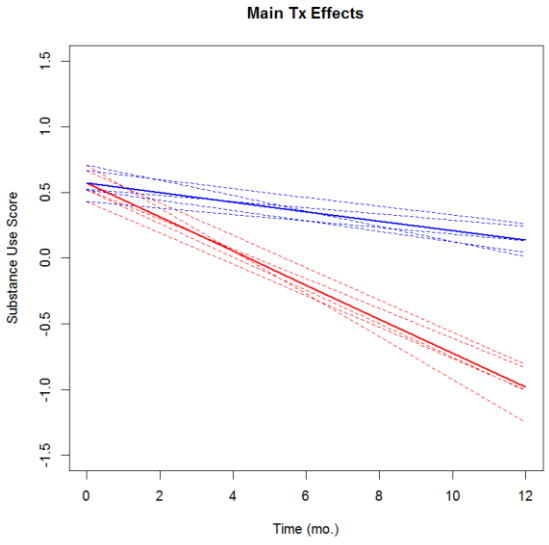


Figure 6.4 Mean Substance Use Factor Scores

Table 6.6 MDFT Slope Effects and Standard Errors

Effect (Std.)	Calibration					
	1	2	3	4	5	Combined
Overall/main	-0.09** (0.026)	-0.1** (0.024)	-0.09** (0.024)	-0.089** (0.026)	-0.105** (0.026)	-0.0986** (0.0189)
Ethnicity						
African Am.	-0.09** (0.020)	-0.1** (0.019)	-0.096** (0.019)	-0.095** (0.02)	-0.109** (0.021)	-0.0986** (0.0171)
Hispanic	-0.07+ (0.037)	-0.08* (0.035)	-0.071* (0.035)	-0.068+ (0.037)	-0.088* (0.038)	-0.0746* (0.0357)
White	-0.11** (0.036)	-0.12** (0.034)	-0.112** (0.034)	-0.112** (0.036)	-0.13** (0.036)	-0.1168** (0.0345)
Gender						
Male	-0.09** (0.025)	-0.1** (0.024)	-0.093** (0.024)	-0.093** (0.026)	-0.111** (0.026)	-0.0978** (0.0239)
Female	-0.07* (0.032)	-0.08* (0.024)	-0.077* (0.03)	-0.074* (0.032)	-0.084** (0.032)	-0.077* (0.0299)

**p<.01, *p<.05, +p<.10



TAU MDFT------ Calibration ——Combined

Figure 5.6 Main LGC Treatment Effects

The combined estimate from each of the five of the individual estimates for the overall MDFT and among African Americans, Whites and males were significant at the .01 level. The MDFT effect among females was significant at the .05 level in the combined estimate and four of the five calibrations, with one calibration significant at the .01 level. Finally, Hispanics showed signficant MDFT effects at the .05 level for the combined estimate and three of the five calibrations, with two of the calibrations only indicating marginally signficant effects at the .10 level.

Summary

MDFT showed greater reductions in substance use trajectories than the comparison treatments overall and across all subgroups, with the largest magnitude of MDFT benefit observed in Whites. There were slight variations observed among the five LGCs, however, they were mostly in agreement. All of the observed treatment effect differences in this analysis were large, so five calibrations were enough to determine the statistical significance. If weaker treatment effect differences were being assessed, more calibrations would be necessary to determine significance with this method. Furthermore, if analysis was limited to a single calibration, two of the five calibrations would have missed the significant MDFT effects in Hispanics.

Results from this analysis were consistent with previous MDFT research in that the largest magnitude of treatment effects was observed in the white and male subgroups. However, interpretation of the observed difference in magnitude was difficult due to the lack of an explicit clinical meaning of a factor score. The significant treatment effects

observed for Hispanics, African Americans and females, though, did provide evidence that these subgroups would benefit more from MDFT over any of the TAU interventions.

Limitations and Future Research

In attempt to reduce complexity, support stability, and increase interpretability in model estimation, only main effects were tested for DIF. However, important interaction terms, particularly study-covariate interactions could provide valuable insight to how the items are functioning differently across these subgroups. The trends observed among the 5 MNLFA model DIF estimates could not come to any concrete conclusions regarding the pattern of significance, though perhaps using a more conservative significance level, such as .01 instead of .05 as the decision criteria for covariate inclusion would have produced more consistent results. A better understanding of how the items are functioning across subgroups would require a more in depth analysis, as well as an applied look into the clinical meaning and cultural context of how 4 indicators in the item set were used in practice.

To avoid the potential for a confounding effect of study by time points on the LGC parameters, two studies for which data was available were excluded because they did not meet the inclusion criteria of having all four of the baseline, 4 month, 6 month and 12 month measurement occasions. Additionally, a 6-week measurement occasion in the ATM study and a 3 month measurement occasion in the ART study were omitted from analysis. Ideally, all relevant information available would be included.

Also, besides treatment non-baseline and treatment non-baseline to study interactions no other predictors of change over time were included in the MNLFA model

which may occlude the substance use trajectories as factor scores for the last three time points were generated assuming the same mean. For this particular application, with only four time points and the tendency for both treatment effects to be exaggerated at treatment termination and relapses in the months thereafter to be common (Burleson & Kaminer, 2007); a non-baseline indicator was determined appropriate to capture the overall, 'leveled-off' substance use reduction from baseline to a one year follow-up.. Any future research implementing this method should consider including more precise indicators of time.

Chapter 7: Summary and Discussion

Simulation Study

This simulation study showed that the MI combining rules are not an appropriate analytic tool to use in the case of combining multiple calibrations. The MI standard error estimator is too conservative, and the MI degrees of freedom formula calculated artificially high values at low levels of *m*. Furthermore, the performance with respect to power using the MI rules only performed as expected in the scenarios with a medium effect size and large sample size, or with a large effect size and any sample size. The MC formulas, however, showed robust performance over all of the investigated scenarios.

Consistent with research by Wang et al. (2013), results from this study suggested that 20 calibrations are ideal for the MC estimator to produce sufficient results, and after 20 calibrations the relative contribution of an additional calibration to the combined estimator was minimal. However, within 5 calibrations, about half of the relative benefit of implementing a MC approach over a single calibration approach was already observed, along with at least some indication of significance.

Of the investigated data characteristics, the number of time points was the only factor that affected the operating characteristics of an MC estimator, with a higher number of time points reflecting a higher efficiency. The sample size only changed the scales of the performance measures, which were observed through systematic, additive affects that were consistent over all levels of *m*. Lastly, while effect size had no influence

on bias or *MSE*, it was the data factor that had the strongest impact on power. When an effect size is strong, only a few calibrations are necessary to pick up the stastical significance.

As a general suggestion to researches interested in this method, 20 calibrations are ideal, but 5 may be sufficient. It is important to note that simulation studies only provide empirical evidence of an analysis method, under hypothetical scenarios (Burton, 2006). To answer research questions on real data, one must be able to appropriately apply them.

MDFT Application

Results from the MC application suggested that males and females, as well as all African American, White, and Hispanic subgroups, would benefit more from MDFT intervention over any of the active control treatment groups. That is, MDFT reduced the trajectories of substance use involvement at a 1-year follow-up to a significantly greater extent, than any of the TAU groups, both overall and across all gender and ethnicity subgroups. However, this analysis was conducted for illustration purposes only and is a much smaller version of an ongoing project being conducted by Greenbaum et al. (2013). The obstacles encountered in the ongoing project which led to the decision in this study to exclude available data as well as potentially important interaction terms, are bound to be encountered by any researcher embarking on an MC approach in an IDA framework. Preferably, rather than excluding data because it complicates the analysis, efforts should be made to find way to incorporate all available information. Nonetheless, this MDFT analysis did show a successful implementation of a MC approach in an IDA framework,

as well as the risk associated with a single calibration approach of missing the presence of significant treatment effects in subgroups.

Discussion

Linking studies at the primary factor level in an IDA framework is an exciting, powerful extension in pooled analysis techniques. However, the derived commensurate measures entailed in the process are statistical constructions that can be hard to interpret. The loss of explicit clinical definitions in using generated factor scores as commensurate measures complicates the portrayal of results and may reignite statistical verse clinical significance controversies. However, many concepts in psychology and social sciences are indeed abstract constructs that elude objective measurement and cannot be directly observed anyway (Bollen et al., 2002). So psychometric models such as an MNLFA that use information from multiple indicators simultaneously as an item set, while obscuring interpretation actually strengthens the measurement validity of the underlying construct. Although latent variable approaches have been and will likely remain controversial for many years (Bollen et al., 2002), a search conducted by DiStefano et al. (2009) uncovered a total of 229 published application articles spanning a variety of disciplines including education, psychology, public health, and law, that created and used factor scores in subsequent analysis. So it may be that in due course, as methodological aspects are disseminated, technological tutorials are made available and suitable software is made accessible, this practice will become more widely understood and accpeted.

Although the art of meta-analysis is now more than a century old (Sutton and Higgins, 2007), the science of integrating data is still evolving (Hussong et al., 2013). And as the corner stone of any field of scientific inquiry is the pursuit of a body of

cumulative knowledge, the IDA framework is a flexible and energetic response to the recent pushes toward collaborative research efforts (Curran & Hussong, 2009; Hussong et al., 2013). Even though rapid advances in statistical methods may be limited to existing software, the development of the MC approach has been pioneered to overcome these limitations. This simulation has demonstrated the statistical advantages of a MC approach to overcome concerns that such as solution instability, and losses in power and precision that single calibration results are subject to. Therefore, any analysis in which a calibration step is employed, multiple calibrations are suggested.

The implementation of multiple calibration approach can be a very time and labor intensive process when used in the framework of an IDA to estimate a MNLFA model. However, the use of multiple calibrations is also not limited to an IDA application and has the potential to be applied in any analysis in which the correlational structure of a given data set needs to be simplified by removing a level of dependency. The relative gain and cost-benefit relationship of an additional calibration's contribution to an MC estimator, though, will be dependent upon the application and data at hand. In any case, the multiple calibration estimator is superior to a single calibration estimator, and more calibrations are always better. Although the MC is analogous to MI, it requires specific combining rules, which have now been made aware and added to the analysis toolkit available for the interested researcher.

Works Cited

Bauer, D. J., & Hussong, A. M. (2009). Psychometric Approaches for Developing Commensurate Measures across Independent Studies: Traditional and New Models. *Psychological Methods, 14*(2), 101-125.

Burleson, J. A., & Kaminer, Y. (2007). Aftercare for adolescent alcohol use disorder: Feasibility and acceptability of a phone intervention. *American Journal on Addictions*, 16(3), 202-205.

Burton, A., Altman, D. G., Royston, P., Holder, R. L. (2006). The Design of Simulation Studies in Medical Statistics *Statistics in Medicine* 25, 4279-4292.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine*, 21(3), 371-387.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology, 28*(1), 1-9.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology,* 53, 605-634.

Bower, P., Byford, S., Barber, J., Beecham, J., Simpson, S., Friedli, K., . . . Harvey, I. (2003). Meta-analysis of data on costs from trials of counselling in primary care: using individual patient data to overcome sample size limitations in economic analyses. *BMJ, 326*(7401), 1247-1250.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A Brief History of Research Synthesis. *Evaluation & the Health Professions, 25*(1), 12-37.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165-176.

Curran, P. J., & Hussong, A. M. (2009). Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets. *Psychological Methods, 14*(2), 81-100.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology*, 44(2), 365-380.

DiStefano, C., Zhu, M., Mindrila, D. (2009). Understanding and Using Factor Scores: Consiterations for the Applied Reseracher. Pracitcal Assesment, Research & Evaluation 14 (20).

Feingold, Alan. (2009). Effect Sizes for Growth-Modeling Analysis for Controlled Clinical Trilas in the Same Metric as for Classical Analysis. *Psychological Methods*, 14 (1), 43-53.

Fisher, D. J., Copas, A. J., Tierney, J. F., & Parmar, M. K. B. (2011). A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of Clinical Epidemiology, 64*(9), 949-967.

Glass, G. V. (2000, March). *The future of meta-analysis.* Paperpresented at the University of California, Berkeley–Stanford University Colloquium on Meta-Analysis, Department of Psychology, University of California, Berkeley.

Greenbaum, PE, Wang, W, Henderson, CE, Hall, K. (2013, May). *Integrative Data Analysis* (*IDA*)*for Longitudinal Data: Pooled Estimates Based On Multiple Calibrations*? Presented at Society for Prevention Research 2013 Annual Meeting. San Francisco, CA.

Graham, J. W., Olchowski A. E., Gilreath T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science.*,8(3) 206-213.

He, Y. & Raghunathan, T. E. (2012). Multiple imputation using multivariate gh transformations. *Journal of Applied Statistics*, 39 (10), 2177-2198.

Henderson, C**.,** Greeenbaum, P., Wang, W., Hall, K., Kan, L., Dakof, G., & Liddle, H. (2013, June). Moderator effects of gender and ethnicity across Multidimensional Family Therapy randomized controlled trials in community and justice settings: An integrative data analysis. Paper presented at the annual meeting of the College on Problems of Drug Dependence, San Diego, CA.

Hofer, S. M., & Piccinin, A. M. (2009). Integrative Data Analysis Through Coordination of Measurement and Analysis Protocol Across Independent Longitudinal Studies. *Psychological Methods, 14*(2), 150-164.

Hussong, A. M., Curran P. J., Bauer D. J. (2013). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, 9. 61-89.

Kim, E. S., Kwok, O., Yoon, M. (2012). Testing Factorial Invariance in Multilevel Data: A Monte Carlo Study. *Structural Equation Modeling: A multidisciplinary Journal*, 19(2), 250-267.

Leccese, M. Waldron, H. B. (1994). Assessing Adolescent Substacne Use: A critique of Current Measurement Instruments. *Journal of Substance Abuse Treatment* 11 6, 553-563.

Liddle, H. A. & Dakof, G. A. (2002). A randomized controlled trial of intensive outpatient, family based therapy vs. residential drug treatment for co-morbid adolescent drug abusers. *Drug and Alcohol Dependence, 66,* S2-S202.

Liddle, H. A., Dakof, G. A., Turner, R. M., Henderson, C. E., & Greenbaum, P. E. (2008). Treating adolescent drug abuse: A randomized trial comparing multidimensional family therapy and cognitive behavior therapy. *Addiction, 103,* 1660-1670.

Liddle, H. A., Rowe, C.L., Dakof, G.A., Henderson, C.E., & Greenbaum, P.E. (2009). Multidimensional family therapy for young adolescent substance abuse: Twelve-month outcomes of a randomized controlled trial. *Journal of Consulting and Clinical Psychology, 77,* 12-25.

Merkle, E. C. (2011) A Comparison of Imputation Methods for Bayesian Factor Analysis Models. *Journal of Educational and Behavioral Statistics* 36 (2), 257-266.

Muthén, L. & Muthén, B. (1998-2011). Mplus user's guide [Computer software]. Los Angeles, CA: Muthén & Muthén.

Nieri, M., Clauser, C., Pagliaro, U., & PiniPrato, G. (2003). Individual patient data: A criterion in grading articles dealing with therapy outcomes. *Journal of Evidence Based Dental Practice, 3,* 122-126.

Riley, R. D., Simmonds, M. C., & Look, M. P. (2007). Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology, 60*(5), 431.e431-431.e412.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. New York: Chapman and Hall.

Siddique, J. Harel, O. and Crespi, C. M. (2012). Addressing Missing Data Mechanism Uncertainty using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial. *Annals of Applied Statistics*, 6 (4), 1814-1837.

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials, 2*, 209-217.

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions, 25*(1), 76-97.

Sutton A. J., Higgins J. P. T. (2008). Recent Developments in Meta-Analysis. *Statistics in Medicine*, 27, 625-650.

Wang, W, Greenbaum, PE, Henderson, CE, Hall, K. (2013, May). *Multiple Calibrations in Integrative Data Analysis: When Needed, How Many, and How to Combine*? Presented at Society for Prevention Research 2013 Annual Meeting. San Francisco, CA.

Appendix A: Simulation Study

Table A.1 Generated Slopes

β_1	D=0.3	D=0.6	D=0.9
T=4	0.1061	0.2121	0.3182
T=8	0.053	0.1061	0.1591
T=12	0.0354	0.0707	0.1061

Table A.2 "True" Slopes

β_1	D=.3	D=.6	D=.9
T=4	0.1079	0.2135	0.3188
T=8	0.0533	0.1063	0.1594
T=12	0.0354	0.0708	0.1061

Table A.3 Population Seeds

	Seeds		
POP	α	β	E
1	5001	7001	9001
2	5002	7002	9002
3	5003	7003	9003
4	5004	7004	9004
5	5005	7005	9005
6	5006	7006	9006
7	5007	7007	9007
8	5008	7008	9008
9	5009	7009	9009

Table A.4 Replication and Calibration Seeds

					Seeds	
RUN	POP	*D*	*T*	*N*	R sets	C sets
1_1	1	0.3	4	100	55011	77011
1_2	1	0.3	4	200	55012	77012
1_3	1	0.3	4	500	55013	77013
2_1	2	0.3	8	100	55021	77021
2_2	2	0.3	8	200	55022	77022
2_3	2	0.3	8	500	55023	77023
3_1	3	0.3	12	100	55031	77031
3_2	3	0.3	12	200	55032	77032
3_3	3	0.3	12	500	55033	77033
4_1	4	0.6	4	100	55041	77041
4_2	4	0.6	4	200	55042	77042
4_2	4	0.6	4	500	55043	77043
5_1	5	0.6	8	100	55051	77051
5_2	5	0.6	8	200	55052	77052
5_3	5	0.6	8	500	55053	77053
6_1	6	0.6	12	100	55061	77061
6_2	6	0.6	12	200	55062	77062
6_3	6	0.6	12	500	55063	77063
7_1	7	0.9	4	100	55071	77071
7_2	7	0.9	4	200	55072	77072
7_3	7	0.9	4	500	55073	77073
8_1	8	0.9	8	100	55081	77081
8_2	8	0.9	8	200	55082	77082
8_3	8	0.9	8	500	55083	77083
9_1	9	0.9	12	100	55081	77081
9_2	9	0.9	12	200	55082	77082
9_3	9	0.9	12	500	55083	77083

Table A.5 Small Effect Size Degrees of Freedom

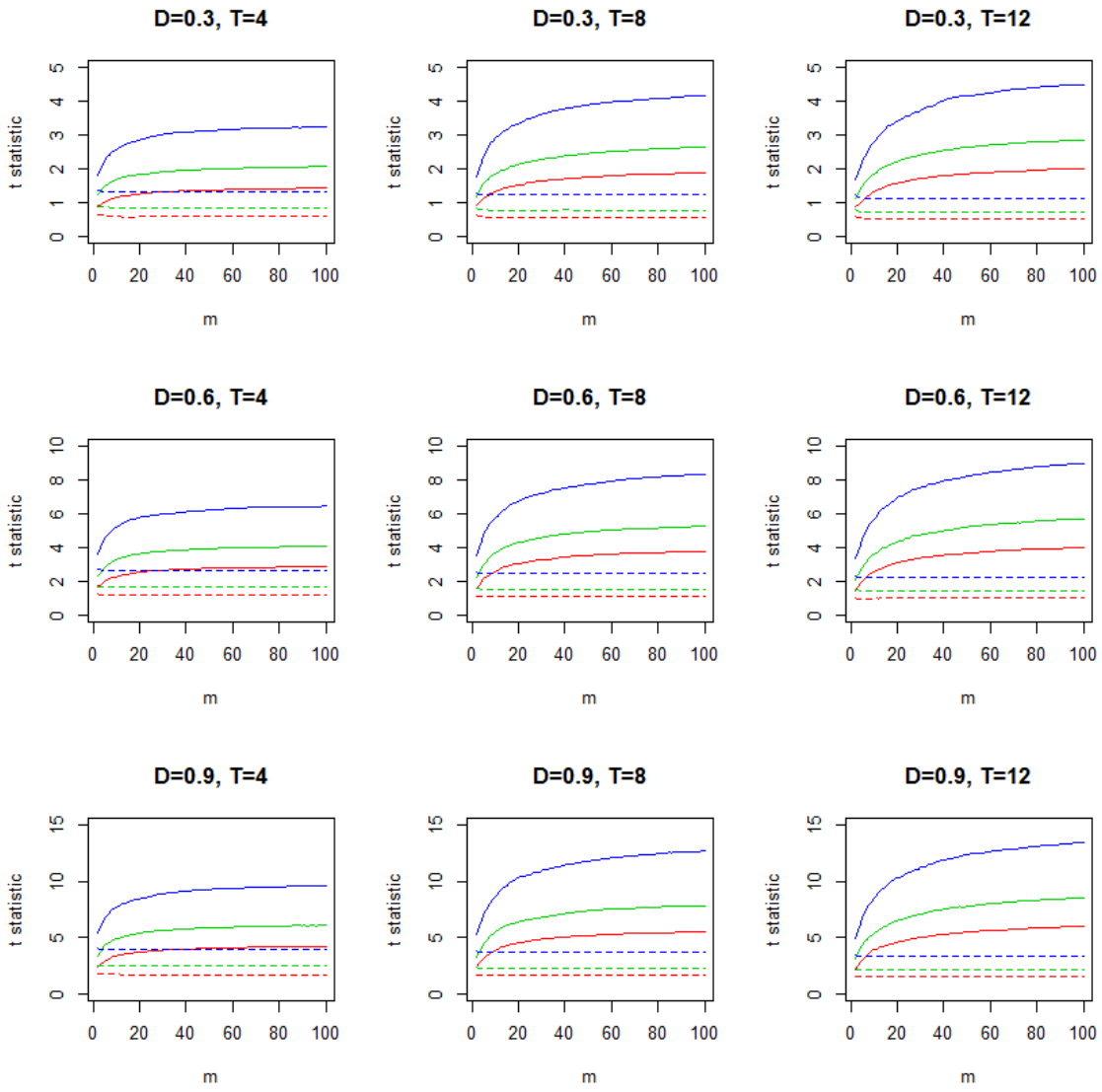
	T=4		*T*=8		*T*=12	
	(MI)	(MC)	(MI)	(MC)	(MI)	(MC)
(*N*=100)						
100	453	9775	423	9788	414	9792
80	362	7815	339	7828	333	7832
40	182	3895	168	3908	165	3912
20	92	1935	87	1948	85	1952
10	53	955	45	968	49	1199
5	48	465	43	799	89	1199
3	1660	399	395	799	14557	1199
(*N*=200)						
100	456	19750	424	19775	418	19783
80	364	15790	338	15815	333	15823
40	181	7870	172	7895	166	7903
20	91	3910	87	3935	84	3943
10	49	1930	48	1955	45	2399
5	186	940	39	1599	36	2399
3	873	799	6169	1599	1027	2399
(*N*=500)						
100	457	49675	429	49738	417	49758
80	365	39715	344	39778	332	39798
40	184	19795	174	19858	165	19878
20	94	9835	91	9898	86	9918
10	52	4855	51	4918	51	5999
5	53	2365	55	3999	231	5999
3	856	1999	17922	3999	1046	5999

Table A.6 Medium Effect Size Degrees of Freedom

	T=4		*T*=8		*T*=12	
	(MI)	(MC)	(MI)	(MC)	(MI)	(MC)
(*N*=100)						
100	459	9775	426	9788	416	9792
80	367	7815	340	7828	332	7832
40	183	3895	169	3908	166	3912
20	91	1935	85	1948	84	1952
10	52	955	48	968	47	1199
5	42	465	79	799	50	1199
3	4285	399	539	799	878	1199
(*N*=200)						
100	450	19750	423	19775	417	19783
80	361	15790	338	15815	335	15823
40	181	7870	168	7895	170	7903
20	96	3910	84	3935	85	3943
10	60	1930	46	1955	48	2399
5	93	940	86	1599	45	2399
3	8910	799	4222	1599	10086	2399
(*N*=500)						
100	452	49675	430	49738	415	49758
80	364	39715	346	39778	332	39798
40	184	19795	177	19858	167	19878
20	95	9835	91	9898	85	9918
10	53	4855	55	4918	48	5999
5	63	2365	52	3999	41	5999
3	114676	1999	86578	3999	4306	5999

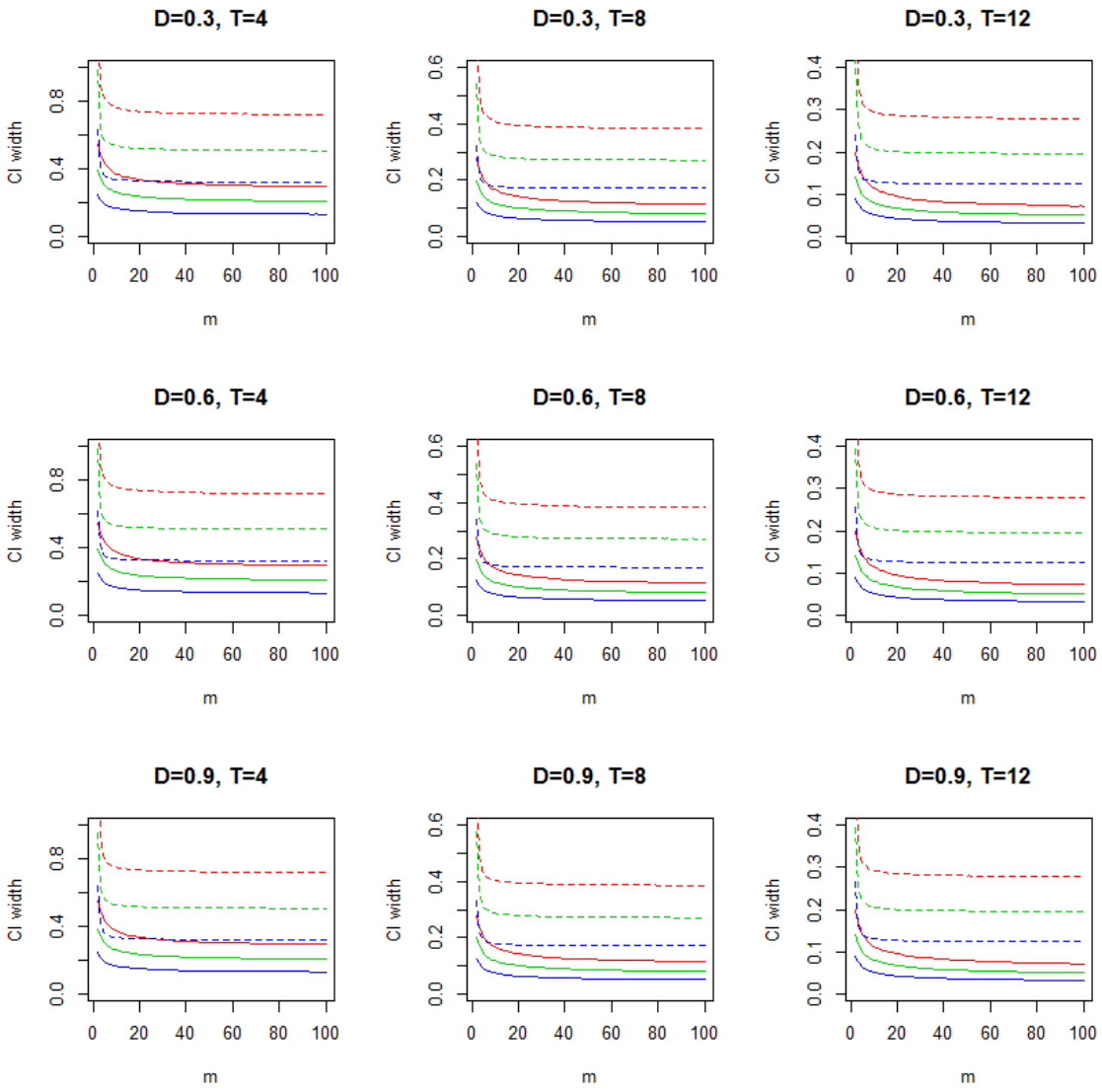
Table A.7 Large Effect Size Degrees of Freedom

	T=4		*T*=8		*T*=12	
	(MI)	(MC)	(MI)	(MC)	(MI)	(MC)
(*N*=100)						
100	460	9775	424	9788	416	9792
80	370	7815	339	7828	331	7832
40	185	3895	169	3908	166	3912
20	98	1935	87	1948	87	1952
10	56	955	47	968	49	1199
5	297	465	69	799	65	1199
3	213	399	1912	799	2728	1199
(*N*=200)						
100	455	19750	424	19775	415	19783
80	365	15790	338	15815	333	15823
40	183	7870	168	7895	167	7903
20	97	3910	86	3935	88	3943
10	55	1930	51	1955	48	2399
5	73	940	57	1599	38	2399
3	6549	799	430	1599	658	2399
(*N*=500)						
100	457	49675	418	49738	415	49758
80	367	39715	335	39778	335	39798
40	187	19795	170	19858	168	19878
20	97	9835	87	9898	84	9918
10	53	4855	51	4918	46	5999
5	55	2365	51	3999	139	5999
3	455	1999	10337868	3999	3001	5999



N=100 *N*=200 *N*=500 —— MC ------ MI

Figure A.1 T-Statistics



N=100 *N*=200 *N*=500 —— MC ------ MI

Figure A.2 Confidence Interval Widths

Appendix B: MDFT Application

Table B.1 Frequency of Outcome Measures by Study

Time	Baseline				4 months				6 months				12 months			
Indicator	AXI	PEI	TLFB	USS	AXI	PEI	TLFB	USS	AXI	PEI	TLFB	USS	AXI	PEI	TLFB	USS
TEM	0	208	223	0	0	110	124	0	0	109	120	0	0	128	136	0
ART	106	107	107	98	101	101	103	95	77	77	102	72	102	102	102	86
ATM	70	0	70	54	58	0	69	53	66	0	68	58	66	0	69	62
Total	176	315	400	152	159	211	296	148	143	186	290	130	168	230	307	148

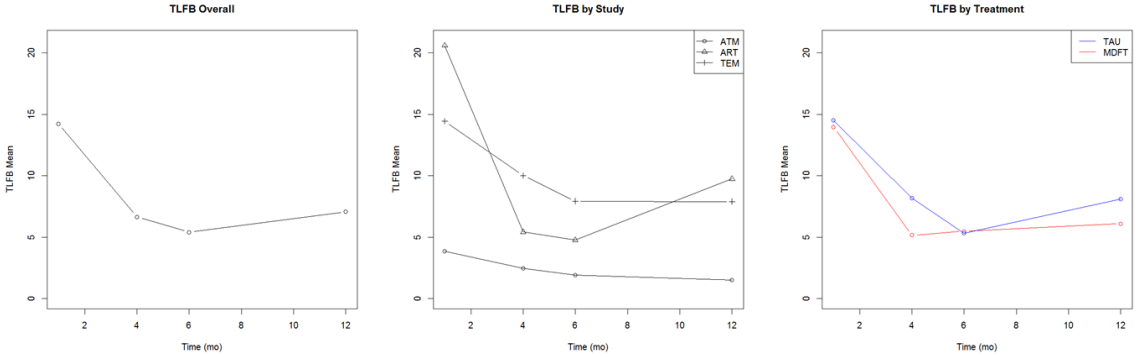


Figure B.1 TLFB Means by Time

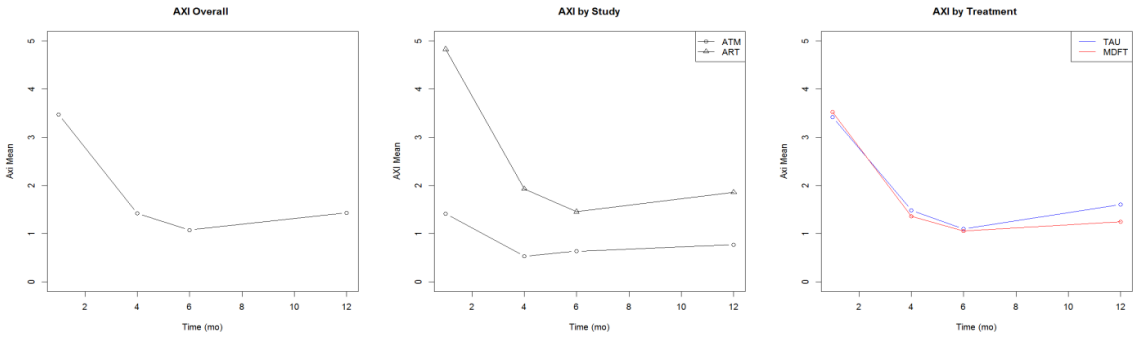


Figure B.2 AXI Means by Time

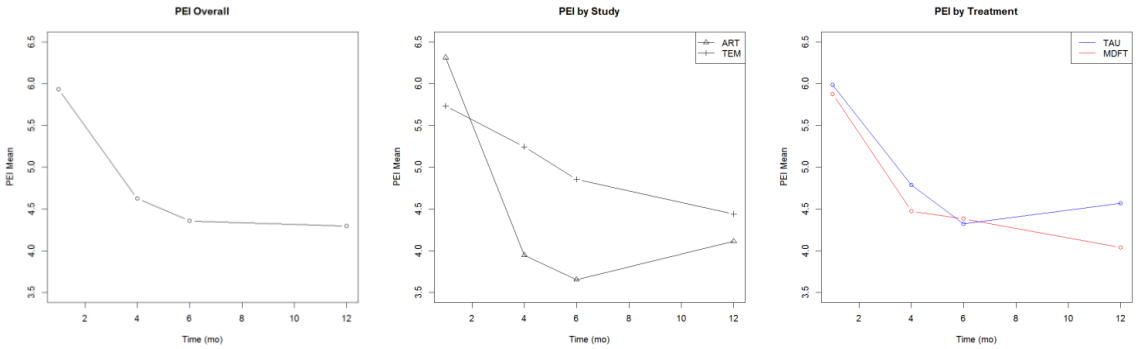


Figure B.3 PEI Means by Time

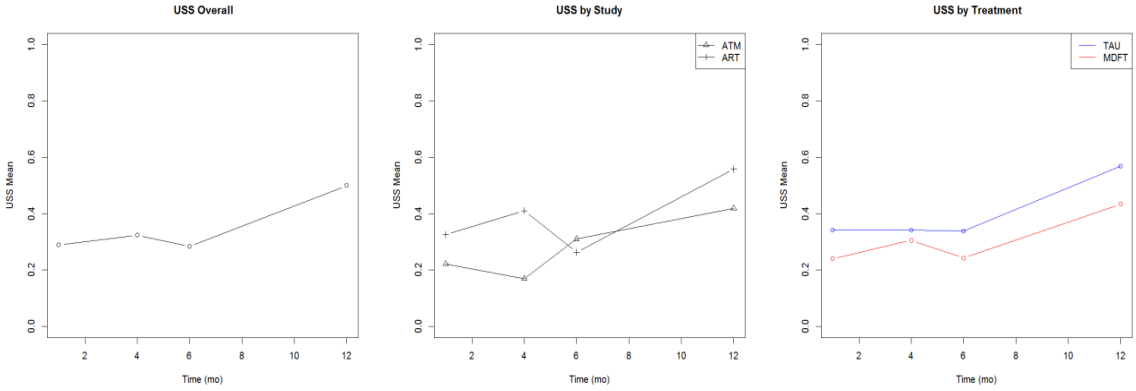


Figure B.4 USS Means by Time

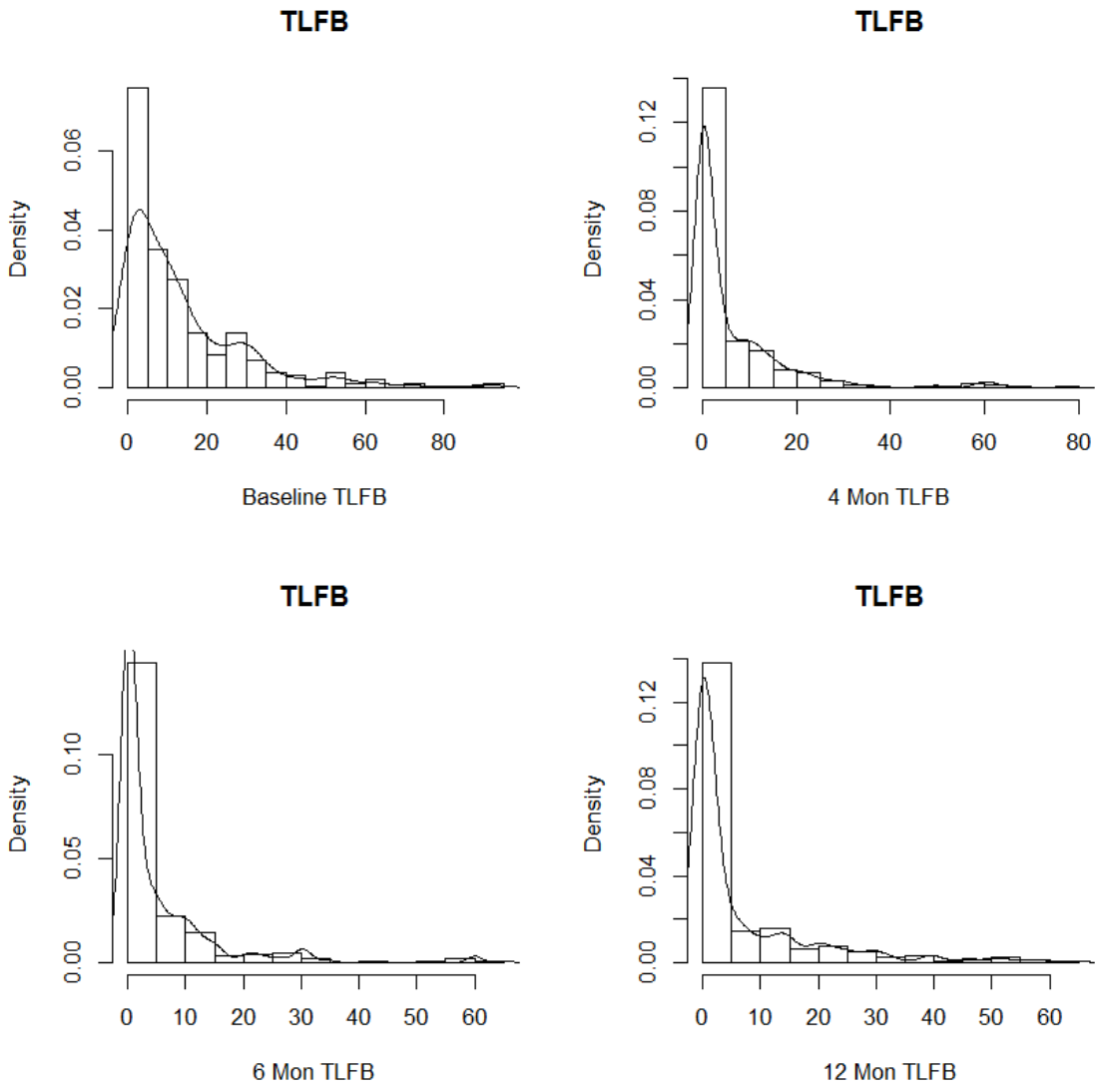


Figure B.5 TLFB Distributions

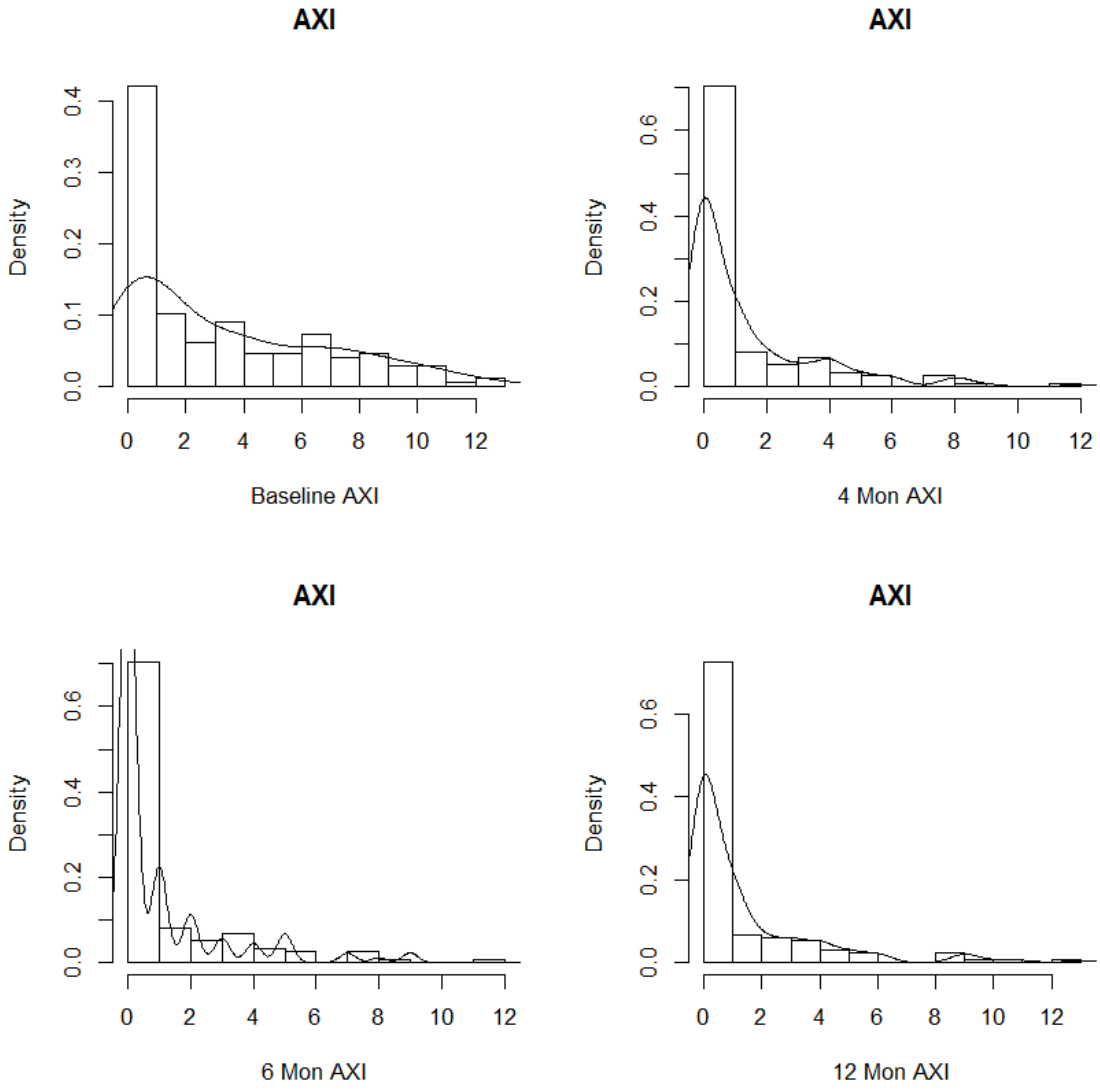


Figure B.6 AXI Distributions

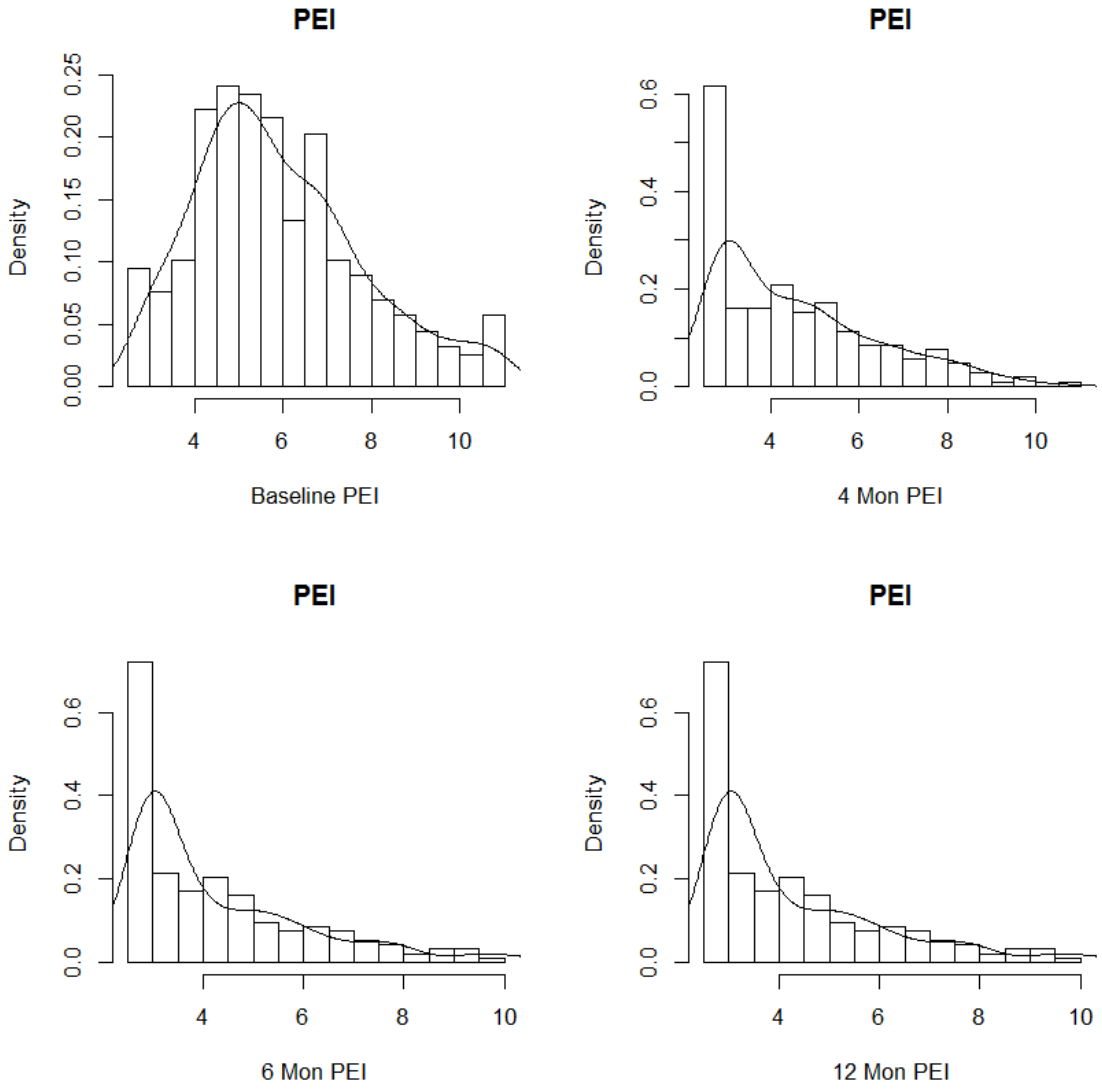


Figure B.7 PEI Distributions

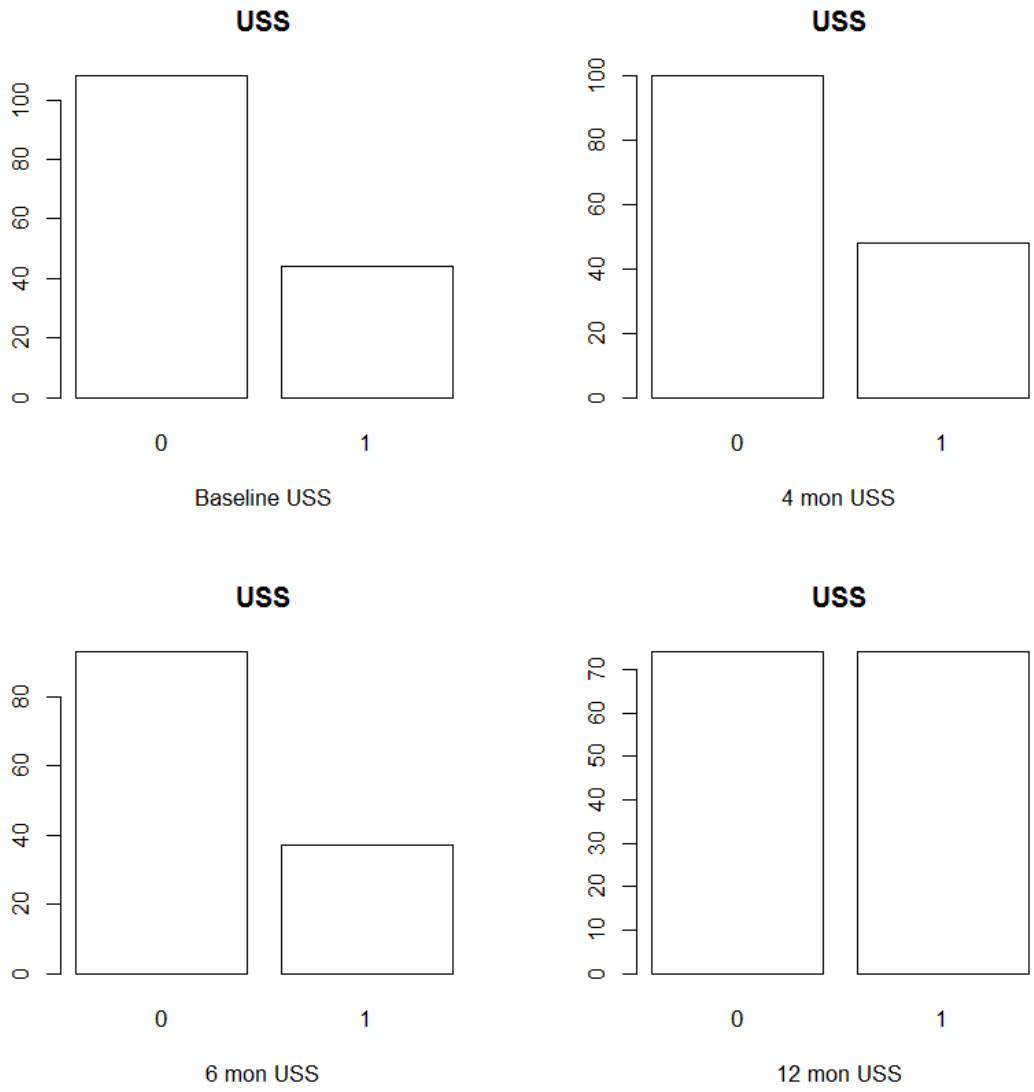


Figure B.8 USS Distributions